

A novel method for crosstalk analysis of biological networks: improving accuracy of pathway annotation

Christoph Ogris^{1,*}, Dimitri Guala¹, Thomas Helleday² and Erik L. L. Sonnhammer¹

¹Stockholm Bioinformatics Center, Department of Biochemistry and Biophysics, Stockholm University, Science for Life Laboratory, Box 1031, 17121 Solna, Sweden and ²Division of Translational Medicine and Chemical Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Science for Life Laboratory, Box 1031, 17121 Solna, Sweden

Received June 3, 2016; Revised August 17, 2016; Accepted August 23, 2016

ABSTRACT

Analyzing gene expression patterns is a mainstay to gain functional insights of biological systems. A plethora of tools exist to identify significant enrichment of pathways for a set of differentially expressed genes. Most tools analyze gene overlap between gene sets and are therefore severely hampered by the current state of pathway annotation, yet at the same time they run a high risk of false assignments. A way to improve both true positive and false positive rates (FPRs) is to use a functional association network and instead look for enrichment of network connections between gene sets. We present a new network crosstalk analysis method BinoX that determines the statistical significance of network link enrichment or depletion between gene sets, using the binomial distribution. This is a much more appropriate statistical model than previous methods have employed, and as a result BinoX yields substantially better true positive and FPRs than was possible before. A number of benchmarks were performed to assess the accuracy of BinoX and competing methods. We demonstrate examples of how BinoX finds many biologically meaningful pathway annotations for gene sets from cancer and other diseases, which are not found by other methods. BinoX is available at <http://sonnhammer.org/BinoX>.

INTRODUCTION

Functional genomics techniques are routinely used to characterize gene expression patterns that result from a particular biological condition. Such data can be used to determine which genes are differentially expressed between e.g. a disease and a normal state. It is however much less trivial to understand how the altered gene expression pattern reflects the altered state of the system. This requires both

knowledge of how genes are organized into functional modules such as pathways or complexes, and a sound strategy to project experimental data onto this knowledge. Ideally, the method should have negligible false positive and false negative rates (i.e. high precision and recall). Designing methods for detecting activated pathways has been the target of many studies in the past (1) but is still an ongoing challenge.

Most current methods determine pathway activation by the statistical significance of the overlap between the differentially expressed genes and the genes within a pathway. This is referred to as Gene Enrichment Analysis (GEA) and typically uses a hypothesis test of the gene overlap based on set theory (2). More advanced Functional Class Scoring algorithms (FCS) like Gene Set Enrichment Analysis can improve the results by using the gene expression level as additional information (3). A major drawback of GEA and FCS methods is the fact that our knowledge of pathways is highly incomplete, which means that the overlap with known pathways is often very small, often resulting in a large number of false negatives (i.e. low coverage). Another issue is that their statistical assumptions require the importance of all genes to be equal and independent which is in conflict with the organization of complex biological systems (3). Pathway topology methods improve the situation somewhat by using known interactions between genes within a pathway as prior knowledge to infer whether it is activated or not (1). This however does not increase the overlap.

Pathway analysis can be improved by using genome-wide functional association networks, like FunCoup (4,5) or STRING (6) as additional evidence. A simple way is to use networks is to apply GEA to a gene set that has been extended with adjacent networks genes. This is e.g. done in FunCoup, STRING and (4,6–8). Network extension approaches might increase the gene overlap but are still subject to the same drawbacks as other gene overlap based methods. A more sophisticated network-based approach is to analyze the network crosstalk (i.e. links) between a query gene set and a pathway, instead of the gene overlap. Here, one assumes a pathway to be activated if a significant enrichment of crosstalk is found (9–12).

*To whom correspondence should be addressed. Institution: Tel: +46 0 70 558 6395; Email: christoph.ogris@scilifelab.se

The fundamental assumption of network-based pathway analysis is that the network consists of functional associations between proteins of the type that may occur within a pathway. The accuracy of these methods depends mainly on two factors. First, the quality of the network—if it has low coverage or poor biological relevance it will not give enough statistical power. Second, the suitability of the statistical model, meaning how well the method can estimate a correct statistical model based on the network, to distinguish spurious from biologically relevant observations. Some approaches assume a normal, Gaussian behavior of crosstalk between gene sets (11), but it has been shown that this only holds if the gene groups and the crosstalk within the network fulfill certain criteria (10). The online pathway annotation method called EnrichNet estimates a distance measure between gene sets via random walk with restart instead of randomizing the whole network (12). The method further relates the raw score to a background model based on all pathways in the database, generating a normalized score between 0 and infinity, yet it does not assess statistical significance of the crosstalk enrichment.

Here we present BinoX, a new network-based algorithm which substantially improves crosstalk significance assessment with a new statistical model. Our benchmarks show that BinoX has considerably better true positive and false positive rates (FPRs) than competing methods such as GEA, NEA (11) and CrossTalkZ (10). It vastly outperforms other network-based approaches in terms of compute time. BinoX can also work with a pre-randomized network which further reduces the compute time and makes it suitable for big data processing.

MATERIALS AND METHODS

Our novel network-based method BinoX was designed for accurate high-throughput analysis of gene set crosstalk, specifically aimed at pathway analysis. It relies on comprehensive functional association networks that combine several different evidence types and represent a general view of gene/protein interactions. The crosstalk between two gene sets, e.g. a pathway and gene signature, is defined as the network connection degree between them, i.e. the number of gene associations. The crosstalk is statistically significant if it is not expected from a random model of the network. Because this random model is unknown we use the principle of Monte-Carlo importance sampling (13) to approximate its parameters. By default, BinoX computes each draw through randomizing the network based on the Link Assignment and Second Order Conservation method (10), which preserves the original second order topological properties and the network degree distribution within the random model. Once the parameters of the random model are computed, it can be used to estimate the probability of an observed crosstalk. In some cases the random model can be approximated with a normal distribution, which has been used in other methods (10,11). The big drawback of using a normal approximation is that one has to discard all cases where the random model does not follow a normal distribution.

Since all network connections are binary, we assume that the crosstalk k between two gene sets follows a binomial distribution $P(k) \sim \text{Bin}(n_k, p_k)$. Here n_k is the number of maximum possible connections between the gene sets, and the probability p_k denotes how likely it is to observe the measured crosstalk. The parameter p_k is approximated through drawing random networks and analyzing the crosstalk within each draw. This Monte-Carlo importance sampling procedure assures the approximation of a null hypothesis and that the results will be more accurate with more random draws. If the expected crosstalk is smaller than the observed, we calculate the P -value for the upper tail, representing enrichment, whereas the opposite can be seen as depletion and is determined based on the lower tail of the distribution. Significant depletion can be interpreted as evidence for the lack of functional association between two gene sets, whereas insignificant crosstalk means lack of evidence for either enrichment or depletion. If BinoX is determining the significance of several crosstalks at once, it applies the Benjamini–Hochberg false discovery rate-correction (FDR) (14) to correct for multiple hypothesis testing statistics. See Figure 1 for an overview of the BinoX algorithm.

Mathematical concepts of BinoX

Assume an undirected genome wide functional association network with r genes. Those genes can be seen as set $V = \{v_1, \dots, v_r\}$. All possible connections $V \times V$ can be represented in a matrix X where

$$x_{ij} = 1 \text{ if } v_i \text{ connected to } v_j \text{ and } i \neq j, \text{ otherwise } x_{ij} = 0$$

Now consider two sets of genes $\{A, B\} \subseteq V$ where the number of connections k between A and B within the network X can be expressed by

$$k = \sum_{i \in A} \sum_{j \in B} x_{ij}$$

Since X is based on binomial relations, k following the binomial distribution $\text{Bin}(n, p)$, where n is the maximum possible connections between A and B and p the probability of observing k . The parameters p and n evolve from a mixture of unknown binomial distributions, hence we can not estimate them directly to approximate a proper null hypothesis to compute the statistical significance of k . Instead we use an alternative hypothesis H_1 to test how likely it is that one observes k by chance. For this alternative distribution we evaluate N shuffled networks, X'_N and define the parameters n' and p' as

$$n' = \min \left\{ |A||B|, \sum_{i \in A} \sum_{l \in V} x_{il}, \sum_{j \in B} \sum_{l \in V} x_{lj} \right\}$$

$$p' = \frac{E(K')}{n'} \text{ where } E(K') = \frac{1}{N} \sum_m k'_m$$

By using Monte-Carlo sampling we estimate $E(K')$ as an approximation for the expected number of connections between $\{A, B\}$ within a random environment X'_N . Here the law of large numbers holds so if $N \rightarrow \infty$, $\text{var}(K') \rightarrow 0$ and $E(K') \rightarrow K'$. We now calculate if A and B are depleted or

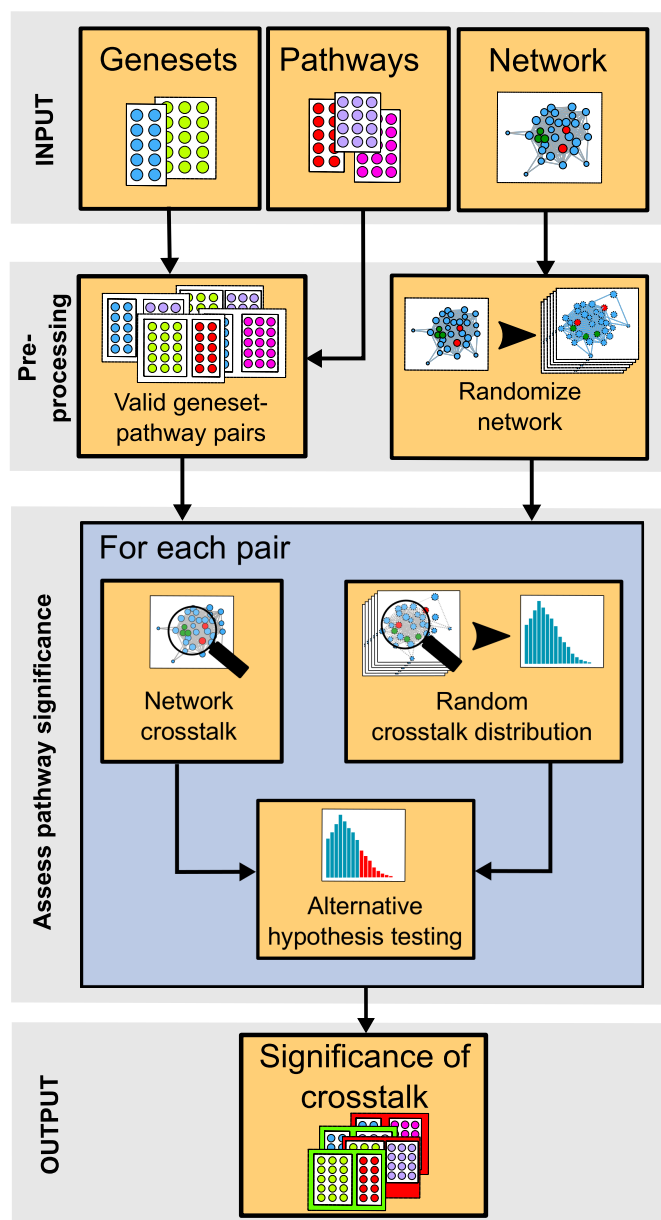


Figure 1. Workflow of BinoX. The algorithm is designed to assess the significance of crosstalk between multiple gene signatures and pathways using a genome wide functional association network as evidence. During preprocessing, BinoX validates gene set-pathway combinations by checking if any crosstalk exists. A random network model is generated based on the original network via Monte-Carlo simulation. The crosstalk significance between each gene set-pathway combination is estimated from the alternative (random) distribution (blue bars) and the observed original crosstalk (red bars). Finally the program summarizes the results and applies the Benjamini–Hochberg procedure to correct for multiple testing to distinguish between significant (green) and insignificant (red) crosstalk.

enriched by testing if k has a significance level lower than α by

$$P(\text{reject } H_1, \text{ crosstalk between } A \text{ and } B \text{ is enriched}) = P(k \geq E(k')) \leq \alpha$$

$$P(\text{reject } H_1, \text{ crosstalk between } A \text{ and } B \text{ is depleted}) = P(k < E(k')) \leq \alpha$$

Stability and robustness

Since the parameter selection of the statistical model is based on Monte-Carlo importance sampling, one can use the concept of the law of large numbers to test for stability and robustness. The idea is that an increased number of random network draws, N , improves our approximation of the unknown random network parameters, which in turn leads to a smaller variance, $\text{var}(K')$, of the estimated results. To test this we ran BinoX with different numbers of random network draws. We used the human FunCoup 3.0 (Schmitt *et al.* (4)) network and estimated in each run the crosstalk between 250 and 1000 randomly generated gene sets. After calculating the relative change of the results we were able to determine an uncertainty error distribution for each setting (see Supplementary Figure S1a). If $N \rightarrow \infty$, $\text{var}(K') \rightarrow 0$ and the Laplace shaped density functions approximates a Dirac delta function resulting in an error of uncertainty of 0. As expected, the variance of the results decreases with increasing number of draws, illustrated in Supplementary Figure S1b. The results based on 10, 150 and 400 random network draws produced a variance of ~ 34.5 , ~ 10 and $\sim 6.2\%$ respectively. This shows that the estimation procedure of BinoX is converging already at a relatively low number of draws indicating that our statistical model is very well suited for the task of describing crosstalk in a random network.

Typical gene overlap

Benchmark datasets can have an arbitrary amount of overlap. However, GEA can only detect pathway enrichment if a gene overlap exists, yet if the overlap is large it will always detect it. To make our benchmark datasets as realistic as possible, we therefore chose a gene overlap level that mimics the gene overlap between all KEGG and MSigDB gene sets, see Supplementary Figure S2.

True positives

To generate a representative test case scenario for evaluating the true positive rate (TPR) we used 288 human KEGG pathways and divided them into two parts with a typical amount of gene overlap. The overlap should simulate a real world example and is needed to calculate enrichment with GEA. To simulate a typical overlap we approximated a distribution representing the overlaps between all KEGG and MSigDB gene sets. As a result 5.6% of the bisected pathways have an overlap of at least one gene. The median value of all overlaps is four genes.

False positives

To test false positives we used the 288 bisected KEGG pathways and randomized them preserving their size and connection degree within the network. To preserve the connection degree distribution of the original bisected pathways, each gene was replaced by a gene having a similar degree in the biological network. Genes present in the other half are allowed to be drawn to obtain a typical overlap to enable detection by GEA. In this dataset, 6.6% of the generated gene set pairs have a gene overlap and these have a mean

overlap of 4.6 genes (overall the mean overlap is 0.3). After swapping the genes, the resulting random bisected pathways have network connection properties close to the original but without biological meaning. Therefore a detected enrichment between the pathway halves is considered as a false positive.

Genome wide functional association network

The human FunCoup 3.0 (4,5) database integrates nine different evidence types, mRNA co-expression, protein–protein interaction, phylogenetic profile similarity, shared transcription factor binding, subcellular co-localization, co-miRNA regulation by shared miRNA targeting, domain interactions, protein co-expression and genetic interaction profile similarity. The FunCoup 3.0 network possesses scale free properties, containing the highest connected node with 4069 links and the smallest with 1. To evaluate the speed we used the human FunCoup 3.0 network at different confidence score cutoffs, see Supplementary Table S1. All other benchmarks and tests are based on the human FunCoup v3.0 network using all links with a confidence score 0.75 (1 123 873 links, 12 391 nodes).

Data access

The 2392 MSigDB gene sets were taken from the C2.CGP (chemical and genetic perturbations) v3.0 collection, downloaded from http://software.broadinstitute.org/gsea/msigdb/download_file.jsp filePath=/resources/msigdb/3.0/c2.cgp.v3.0.symbols.gmt. The 288 human pathways were gathered from KEGG Release 70.1 via <http://www.kegg.jp/kegg/rest/keggapi.html>. Furthermore we used the human genome wide functional association network from FunCoup v3.0 build 2014-02 available at <http://funcoup.sbc.su.se/downloads/>.

Used software

NEA http://research.scilifelab.se/andrej_alexeyenko/downloads.html, CrossTalkZ v1.3.3 <http://sonnhammer.sbc.su.se/download/software/CrossTalkZ/>, BinoX v0.9.3 <http://sonnhammer.org/BinoX>

RESULTS

Benchmarking true positive rate

Functionally associated gene sets are likely to have significant crosstalk between them. This implies that functional gene sets, for instance pathways, are likely to have significant intra-crosstalk as well. We thus reasoned that a benchmark of true positives may be constructed by splitting known pathways into parts and measuring a method's ability to re-connect these parts.

A benchmark dataset was constructed from 288 human pathways from KEGG (15) which were divided into two parts with typical amounts of gene overlap (see 'Materials and Methods' section), otherwise GEA would find no true positives. To make it more challenging we also created a dataset of the 25% smallest pathways only.

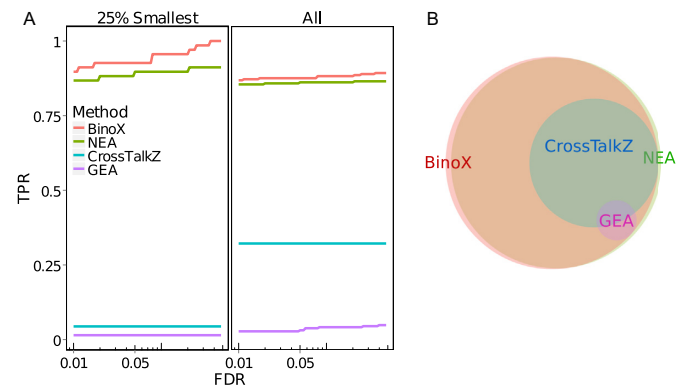


Figure 2. True positive rate (TPR) benchmark of BinoX, NEA, CrossTalkZ and GEA. The test was performed using KEGG pathways separated into two parts with typical gene overlap. A true positive is defined as two parts from the same pathway having significant crosstalk. (A) The right panel shows the TPR for all pathways, and the left panel for the 25% smallest pathways. (B) Venn Diagram showing the overlap of true positives detected at $FDR \leq 0.05$ by BinoX (98.1% of all detected), NEA (96.5% of all detected), CrossTalkZ (36.0% of all detected) and GEA (3.5% of all detected).

The advantage of using a network becomes obvious through a direct comparison between network-based and non-network-based methods. For $FDR \leq 0.05$, GEA restores only 1.4% of the pathways correctly, while BinoX at 87.5% TPR ranks in first place followed by NEA at 86.1% and CrossTalkZ at 32.1%. For the 25% smallest pathways GEA's TPR stays at 1.5%, while BinoX increases to 92.6%, followed by NEA at 89.7% and CrossTalkZ which drops to 4.4% (see Figure 2A). If one looks at the true positives found by all methods at $FDR \leq 0.05$, BinoX identified all true positives found by GEA and CrossTalkZ, and only missed 1.9% which were detected by NEA (Figure 2B). About 3.5% were uniquely identified by BinoX.

Benchmarking false positive rate

To benchmark FPR we generated two datasets, each with 288 random gene sets with properties like size and connection degree distribution chosen to mimic the bisected KEGG pathways above. This ensures that the gene sets lose their biological relevance, implying that any statistically significant crosstalk between gene sets is a false positive. The procedure introduced a typical degree of gene overlap, see Materials and Methods, which enables detection by GEA.

At $FDR \leq 0.05$, BinoX and CrossTalkZ assigned no false positives at all whereas NEA detected 12.5% false positives. The overlap-based GEA assigned 84% of the pairs sharing a gene as significant, which led to an overall FPR of 5.7% (see Figure 3). For the more challenging task of detecting enrichments correctly in the smallest 25% of the gene sets, GEA can't test a single pair because none of them has an overlap. NEA's FPR increased to 23% while BinoX still found no false positives. As the crosstalk between the smallest gene sets is not normally distributed, the normality test in CrossTalkZ failed for all of them, and hence reported no false positives.

Furthermore we combined the FPR with TPR benchmarks in a partial Receiver Operating Characteristics curve,

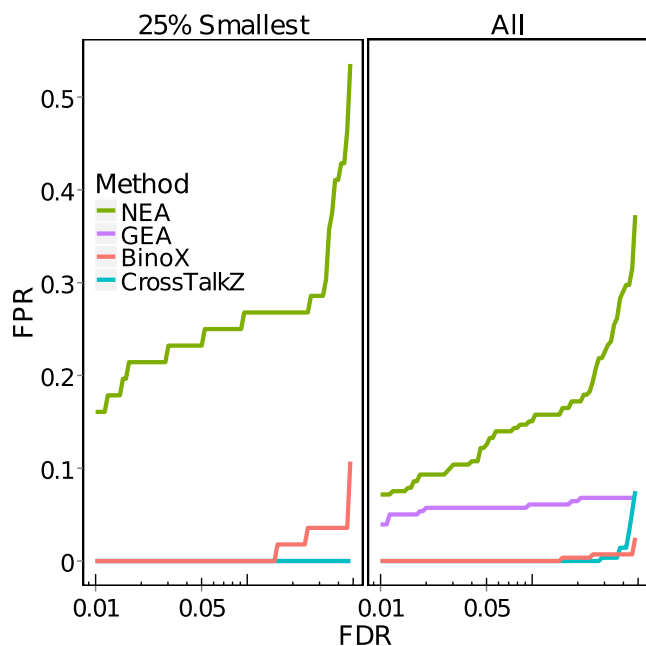


Figure 3. False positive rate benchmark of CrossTalkZ, NEA, GEA and BinoX. We generated random gene sets, mimicking the network structure of bisected KEGG pathways. Two random gene sets having significant crosstalk are defined as a false positive. The right panel shows the FPR for all gene sets, while the left panel is limited to the 25% smallest. As there are no overlapping genes in the latter dataset, GEA is unable to test any associations.

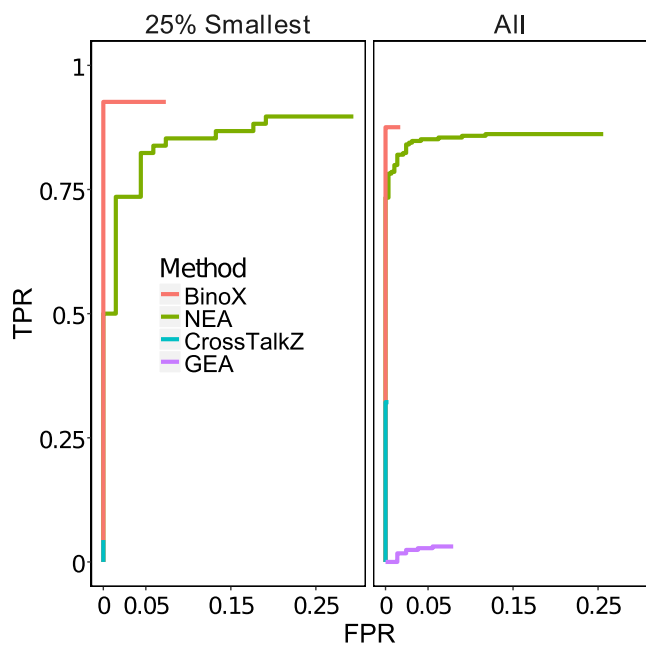


Figure 4. Partial Receiver Operating Characteristics curve, combining TPR and FPR while excluding all insignificant hits ($FDR \geq 0.05$). Here GEA is not visible for the smallest 25% since it detects neither FP nor TP.

excluding all insignificant results at $FDR \leq 0.05$ (see Figure 4). Including all gene sets, BinoX finds 88% of the true positives before including the first false positive. Here NEA follows with 74%, CrosstalkZ with 32% and GEA with 0.3%.

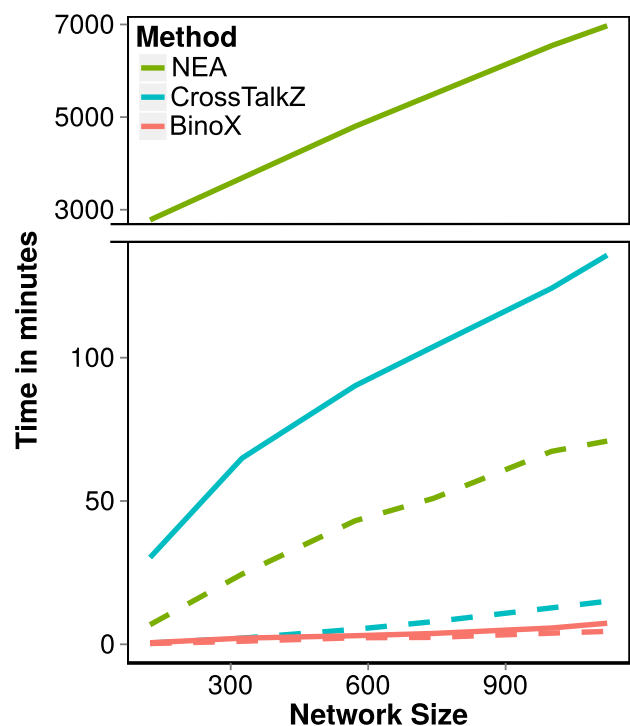


Figure 5. Compute times for CrossTalkZ, NEA and BinoX. Dashed lines: using a gene set of 120 randomly selected genes and calculating the crosstalk to itself for different network sizes. Solid lines: crosstalks between 2392 gene signatures and 288 KEGG pathways. Here NEA is shown with a separate scale. We did not use BinoX's pre-randomization mode which would have skipped the most computationally expensive part. The networks are from FunCoup 3.0 with link confidence 0.75, 0.8, 0.9, 0.9, 0.95, 0.99 and 1 (see Supplementary Table S1).

For the smallest 25% of the gene sets, BinoX improves its false positive free performance to 93% while NEA drops to 50%, CrossTalkZ to 5%, while no hit was reported by GEA.

Benchmarking speed

One of the biggest challenges in network-based approaches is limiting the compute time, to be applicable for analysis of big data. Since the compute time of network-based algorithms depends heavily on the size of the network, we applied a range of cutoffs to the scale-free FunCoup 3.0 network to make six networks of different sizes. For testing speed we designed two scenarios. The first scenario represents a common gene set analysis task where we calculate the intra-crosstalk significance of a gene set with 120 genes. The second scenario tests whether the algorithms can handle a large scale task of running 2392 experimental gene sets in MSigDB (16) against KEGG, which amounts to 691 288 crosstalk assessments.

The first scenario shows that for the smallest network, where BinoX and CrossTalkZ only took a few seconds, NEA ran for ~6 min (see Figure 5). For the largest network, NEA took 71 min, CrossTalkZ 15 min and BinoX only 5 min. This time can be reduced to merely 45 s by setting BinoX's option to use a pre-randomized network as statistical baseline.

In the second scenario, using the smallest network, BinoX took about half a minute, CrossTalkZ 30 min and NEA 46 h. For the largest network, NEA took over 110 h, CrossTalkZ 2.2 h and BinoX 7 min, which is around 1000 times faster than NEA. Using the pre-randomized network this shrinks down to 2.2 min.

Pathway annotation with BinoX

The primary application of BinoX is to efficiently find associations between experimental gene sets and known pathways, in order to obtain functional insights for a given condition. Its superior accuracy offers a high chance of arriving at a correct pathway annotation. To exemplify how BinoX can be used we performed two large-scale analyses: pathway annotation of 2392 experimental gene sets in MSigDB and of 288 disease gene sets. Furthermore we analyzed a single cancer gene set. For all studies we used the FunCoup 3.0 network and generated a statistical model based on 150 network draws. We compared BinoX's results to the most commonly used pathway annotation algorithm, GEA, based on gene overlap as implemented in DAVID (17) and to the network-based CrossTalkZ approach. Due to its excessive compute time and huge FPR we excluded the NEA algorithm.

Pathway annotation of MSigDB. We evaluated crosstalk between the 2392 experimental MSigDB gene sets ('signatures') and KEGG pathways, in total 688 896 signature-pathway comparisons. With FunCoup 3.0, network-based algorithms can assess the crosstalk of 85.5% (510 423) of these. Since GEA is based on gene overlap it can only test 11.4% (68 218) of all cases. At a significance cutoff of $FDR = 0.05$, all three methods detected 173 645 significantly enriched and 63 480 depleted signature-pathway pairs, see Table 1. The enrichments detected by GEA were almost fully covered by BinoX and CrossTalkZ—only 2.3% (4 044) were detected by GEA alone. Given that BinoX and CrossTalkZ had zero false positives in the benchmark above at this FDR level, while GEA had 6%, these are likely to be false positives. 9.4% (16 341) of CrossTalkZ enrichments were not covered by BinoX, BinoX further found 13.1% (22 806) enrichments not covered by the other methods, see Figure 6. Full results are in Supplementary Data 1. At $FDR < 0.001$ BinoX found enriched pathways for 1196 more signatures than GEA (2058 versus 862, see Table 2).

An example of novel pathway annotations found by BinoX is the signature WATANABE_ULCERATIVE_COLITIS_WITH_CANCER_UP. BinoX identified 20 significant KEGG pathways, of which the most significantly enriched was *Extra Cellular Matrix (ECM) Receptor Interaction*, reflecting the fact that Ulcerative Colitis (UC) as well as other Inflammatory Bowel Diseases show organ-specific changes in both the composition and organization of the ECM (18). UC is a digestive disease, resulting in problems with digestion leading to inflammation, which is supported in the BinoX results by the second most significantly enriched pathway *Protein Digestion and Absorption*. Cancer related pathways were also significantly enriched. These include both general pathways like *Pathways in Cancer*, and more specific ones,

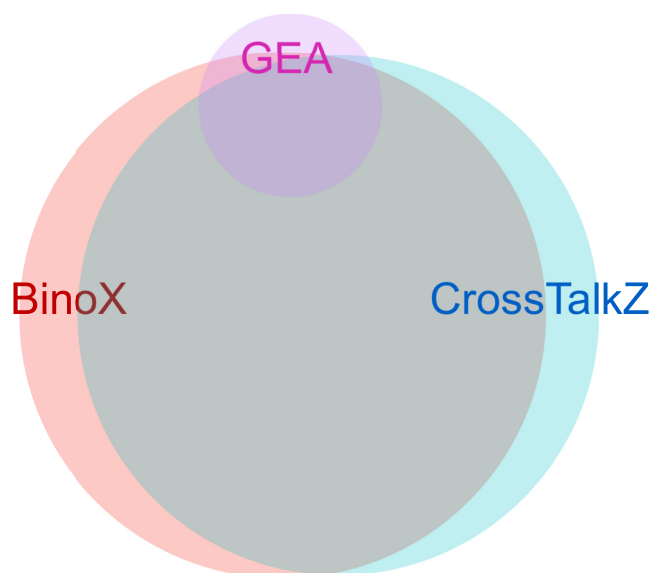


Figure 6. KEGG Pathway annotation for MSigDB. The Venn diagram is based on all significant pairs between 288 KEGG pathways and 2392 MSigDB gene signatures identified at a $FDR \leq 0.05$. BinoX identified 207 600, CrossTalkZ 202 670 significant depletions and enrichments whereas GEA found only 25 106 significant enrichments.

such as *Basal Cell Carcinoma* which is linked to UC via a lack of the same plasma factor (19), and *Proteoglycans in Cancer*, where proteoglycans are both tied to the intestine mucosa and are key to different aspects of cancer biology (20). For this signature, CrossTalkZ found no significant pathways, while GEA only identified *WNT Signaling*.

A further example is the LINDGREN_BLADDER_CANCER_CLUSTER_2A_UP gene set, where BinoX found the *ErbB signaling* pathway at the top of 81 enriched pathways. CrossTalkZ found two significant pathways, and GEA found none. *ErbB signaling* is involved in several aspects of cancer biology, e.g. proliferation, differentiation, cell motility and survival and is a common target in cancer treatment (21,22). Beside several cancer pathways, also pathways in the immune response including *T-cell receptor* and *NK cell-mediated cytotoxicity* were enriched. Additionally, the *Hepatitis B Virus (HBV)* pathway, one of the top enriched, potentially depicts the reactivation of HBV in Bladder cancer patients undergoing chemotherapy (23). The *MicroRNAs in Cancer* pathway was also at the top of the enriched KEGG pathways, and is supported by recent implication of microRNA expression in bladder cancer (24). The *RAS Signaling* pathway has an established role in bladder cancer (25), while the enriched *Estrogen* pathway is suspected to take part in proliferation of urothelial cells in Bladder cancer (26). The pathway *Bacterial Invasion Of Epithelial Cells* was also enriched for this bladder cancer gene set, which may hint to the role of *Bacillus Calmette-Guerin (BCG)* in triggering the immune response in Bladder cancer. BCG has been used as immunotherapy with similar effect sizes as seen in traditional chemotherapy (27).

Table 1. Enrichments and depletions found by GEA, BinoX and CrossTalkZ for different datasets at FDR \leq 0.05

Dataset	Number of (% of union) enriched / depleted pathways			
	GEA	BinoX	CrossTalkZ	Union
Cancer genes	32 (21.8%)/-	147 (100%)/64 (100%)	145 (98.6%)/54 (84.4%)	147/64
Disease genes	7352 (26.9%)/-	24 431 (89.3%)/5810 (86%)	23 590 (86.3%)/6 046 (88.6%)	27 333/6822
MSigDB genes	25 106 (14.6%)/-	153 255 (88.2%)/54 340 (85.6%)	145 020 (83.5%)/57 630 (90.8%)	173 645/63 480

The union represents the amount of significant pathways found by all methods combined.

Table 2. KEGG pathway enrichment or depletion of MSigDB signatures at FDR \leq 0.001

Method	Signatures with enrichments	Mean number of enriched pathways per signature	Signatures with depletions	Mean number of depleted pathways per signature
GEA	862	2.1	0	0
BinoX	2058	40.5	1893	12.1
CrossTalkZ	2054	46.6	1762	13.1

Pathway annotation of cancer genes. We used BinoX to calculate the significance between 288 KEGG pathways and the cancer gene set published by (28). This set contains 412 genes that are mutated and causally implicated in cancer development. BinoX identified 147 enriched and 64 depleted KEGG pathways, while CrossTalkZ identified 145 and 54 respectively. GEA only identified 32 enriched terms as significant at FDR $<$ 0.05 (Supplementary Table S2), all of which were also significantly enriched with the network-based methods, see Table 1. One of the pathways BinoX identified as having a highly significant crosstalk with the cancer gene set was the *Neurotrophin Signaling* pathway. It contains 120 genes and shares only 14 of these with the cancer gene set, resulting in an FDR of 0.07 for GEA. However, in FunCoup the two groups are connected by 888 network links. BinoX estimated that by chance one would observe about 440 links, resulting in a highly significant FDR of 5.8×10^{-80} . It has been shown that neurotrophin activates members of the tumor necrosis factor receptor superfamily (29). Furthermore, GEA missed to identify several other central pathways in cancer, identified by BinoX (and CrossTalkZ) such as the *MAPK Signaling* (30), at FDR = 2.4×10^{-69} and *RAS Signaling*, at FDR = 5.2×10^{-69} . The *MAPK Signaling* pathway regulates multiple cellular functions including proliferation, growth and apoptosis, and is one of the most dysregulated pathways in cancer (31). *RAS* is a sub-pathway to *MAPK*. The two pathways uniquely identified by BinoX, *Gabaergic Synapse* and *Lysine Degradation*, have both been shown to have potential implications in cancer (32,33).

Of the significantly depleted pathways, the lowest FDR (5.2×10^{-208}) was obtained for *Olfactory Transduction* pathway which consists of 294 genes, and had 73 links to the cancer gene set. Given the network and its properties, BinoX calculated an expected number of 713 links. Since *Olfactory Transduction* pathway is part of the sensory system, involving signal transmission to the brain, it makes sense that the cancer gene set is not functionally associated with it. Also other brain related pathways like *Parkinson's Disease* pathway (FDR = 3.1×10^{-17}) and *Alzheimer Disease* pathway (FDR = 1.63×10^{-13}) were depleted, which indicates that these are totally independent from the cancer gene set.

Pathway annotation of diseases. We evaluated crosstalk between 299 diseases that have at least 20 associated genes in the Online Mendelian Inheritance in Man and genome-wide association study databases (34), and KEGG pathways. Full results are in Supplementary Data 2. Of the total 86 112 possible disease-pathway pairs, 28.2% (24 431) and 27.3% (23 593) were statistically enriched (FDR \leq 0.05) in BinoX and CrossTalkZ respectively, while only 9.2% were enriched in GEA (7 952), see Table 1. The overlap between BinoX and CrossTalkZ was 21 525.

For several diseases, BinoX was the only method able to identify several significant crosstalks. One such example is Hypothalamic disease, for which BinoX identified 24 significant pathways. CrossTalkZ and GEA found one pathway each. The second most significantly enriched pathway was *Hedgehog Signaling*, shown to be affected in the Pallister-Hall syndrome, a Hypothalamic disease (35). Additionally, several synaptic pathways (e.g. *Cholinergic*, *Serotonergic*, *Glutamatergic* and *Dopaminergic*) were also enriched for this disease class.

Another example is Vision disorder, where one of the most enriched pathways among the 34 found by BinoX is *Oxidative Phosphorylation* which is directly implicated in Leber optic atrophy (<http://omim.org/entry/535000>). CrossTalkZ and GEA found only one pathway each. Furthermore, BinoX detected *Phototransduction* pathway, shown to be affected in several vision disorders, e.g. Retinitis pigmentosa (36), and Congenital Stationary Night Blindness (37).

DISCUSSION

We have presented a new network-based method for pathway annotation, BinoX and have shown that it offers superior performance compared to existing methods in terms of true positive and FPRs. Compared to other network-based methods it further offers large speed improvements of up to three orders of magnitude. BinoX was designed with pathway annotation in mind but may be employed for any type of gene set analysis as long as a comprehensive functional association network exists for the genes.

Pathway annotation is currently mostly performed with tools that look for gene overlap between experimental gene sets and known pathways. A great number of tools exist that

use variants of this technique. For instance, a recently compiled corpus (2), lists 68 such tools. Because it would be impossible to include all of these in this study, we chose the method employed by one of the most popular pathway annotation tools, DAVID (17). They use the ‘EASE score’ (38) which is a conservative adjustment of Fisher’s exact test. As noted by the authors, the normal Fisher exact test is very sensitive to small pathways and is likely to assign significance to an overlap of only one gene if the pathway is sufficiently small. This problem is reduced with the EASE score by subtracting 1 from the overlap, but as shown in this study the FPR is still very high. In fact, a pathway that shares two genes with a gene set reaches significant enrichment at $P < 0.05$ if they contain less than 33 genes each. A practical problem of GEA is that a given gene set which is not significantly overlapping a KEGG pathway, may be significant, with the same overlap, to the same pathway in e.g. PANTHER if the latter has fewer genes. Because GEA methods only use overlapping genes as evidence, they are limited by the state of current pathway databases. Although GEA could only test 6.6% of the gene set pairs in the false positive benchmark, it assigned 84% of these as significant. With such a high risk of false assignment, it is questionable if GEA is at all suitable for pathway annotation. This is highlighted by the pathway annotation examples where GEA fails to find a large number of relevant pathways. In case of the pathway annotation of MSigDB, 90% of all possible signature-pathway pairs do not share any genes at all, meaning they can not be analyzed by overlap based methods like GEA.

Although CrossTalkZ uses the same association network as BinoX, it often fails because normality in the link distribution is violated. This is witnessed in the benchmarks and examples where BinoX was the only method able to uncover important relationships between MSigDB gene sets and KEGG pathways. NEA uses the same statistical model as CrossTalkZ, i.e. assumes a normal distribution, but does not check for its validity. This results in a higher TPR, but at the expense of a huge FPR. CrossTalkZ on the other hand has a good FPR but at the expense of a very poor TPR. Only BinoX manages to achieve a good balance between these performance measures.

For online pathway analysis of single gene sets, we have set up a web server PathwaX.sbc.su.se (39) which applies the BinoX algorithm to KEGG pathways and FunCoup networks.

Network-based methods can detect both crosstalk enrichment and depletion, whereas gene overlap based methods can only detect enrichment. In what way is significant depletion of a pathway functionally relevant? In our view there are three types of crosstalk: significant enrichment, significant depletion and insignificant crosstalk. Significant crosstalk enrichment, as with gene overlap enrichment, informs about pathway activation. Significant crosstalk depletion means that there is strong evidence that the pathway is not activated. It however does not necessarily mean that the pathway was ‘actively shut down’, only that there is statistical evidence against it being activated, which may originate from a number of different scenarios.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

AstraZeneca [N1372]. Funding for open access charge: Swedish Research Council.

Conflict of interest statement. None declared.

REFERENCES

1. Khatri,P., Sirota,M. and Butte,A.J. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**, e1002375.
2. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
3. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
4. Schmitt,T., Ogris,C. and Sonnhammer,E.L.L. (2014) FunCoup 3.0: database of genome-wide functional coupling networks. *Nucleic Acids Res.*, **42**, D380–D388.
5. Alexeyenko,A. and Sonnhammer,E.L.L. (2009) Global networks of functional coupling in eukaryotes from comprehensive data integration. *Genome Res.*, **19**, 1107–1116.
6. Szklarczyk,D., Franceschini,A., Wyder,S., Forslund,K., Heller,D., Huerta-Cepas,J., Simonovic,M., Roth,A., Santos,A., Tsafou,K.P. *et al.* (2014) STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.
7. Montojo,J., Zuberi,K., Rodriguez,H., Bader,G.D. and Morris,Q. (2014) GeneMANIA: fast gene network construction and function prediction for Cytoscape. *F1000Res*, **3**, 153.
8. Zuberi,K., Franz,M., Rodriguez,H., Montojo,J., Lopes,C.T., Bader,G.D. and Morris,Q. (2013) GeneMANIA prediction server 2013 update. *Nucleic Acids Res.*, **41**, W115–W122.
9. Shojaie,A. and Michailidis,G. (2010) Network enrichment analysis in complex experiments. *Stat. Appl. Genet. Mol. Biol.*, **9**, 1–36.
10. McCormack,T., Frings,O., Alexeyenko,A. and Sonnhammer,E.L.L. (2013) Statistical assessment of crosstalk enrichment between gene groups in biological networks. *PLoS One*, **8**, e54945.
11. Alexeyenko,A., Lee,W., Pernemalm,M., Guegan,J., Dessen,P., Lazar,V., Lehtiö,J. and Pawitan,Y. (2012) Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. *BMC Bioinformatics*, **13**, 226.
12. Glaab,E., Baudot,A., Krasnogor,N., Schneider,R. and Valencia,A. (2012) EnrichNet: network-based gene set enrichment analysis. *Bioinformatics*, **28**, i451–i457.
13. Siegmund,D. (1976) Importance sampling in the Monte Carlo study of sequential tests. *Ann. Stat.*, **4**, 673–684.
14. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Stat. Methodol.*, **57**, 289–300.
15. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
16. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdóttir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.
17. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
18. Järveläinen,H., Sainio,A., Koulou,M., Wight,T.N. and Penttinen,R. (2009) Extracellular matrix molecules: potential targets in pharmacotherapy. *Pharmacol. Rev.*, **61**, 198–223.
19. Klein,A., Storch,B., Meyerson,E., Kadish,U., Kaufmann,H. and Feuerman,E.J. (1982) Basal cell carcinomas, gastrointestinal carcinomas, and ulcerative colitis in relation to the disappearance of a plasma factor. *Cancer Detect. Prev.*, **6**, 527–530.

20. Bandyopadhyay, S.K., Carol, A., Kessler, S.P., Hascall, V.C., Hill, D.R. and Strong, S.A. (2008) Hyaluronan-mediated leukocyte adhesion and dextran sulfate sodium-induced colitis are attenuated in the absence of signal transducer and activator of transcription 1. *Am. J. Pathol.*, **173**, 1361–1368.
21. Gao, H. and Zhang, Z. (2015) Systematic analysis of endometrial cancer-associated hub proteins based on text mining. *Biomed Res. Int.*, **2015**, 615825.
22. Hynes, N.E. and MacDonald, G. (2009) ErbB receptors and signaling pathways in cancer. *Curr. Opin. Cell Biol.*, **21**, 177–184.
23. Higashiyama, H., Harabayashi, T., Shinohara, N., Chuma, M., Hige, S. and Nonomura, K. (2007) Reactivation of hepatitis in a bladder cancer patient receiving chemotherapy. *Int. Urol. Nephrol.*, **39**, 461–463.
24. Yoshino, H., Seki, N., Itesako, T., Chiyomaru, T., Nakagawa, M. and Enokida, H. (2013) Aberrant expression of microRNAs in bladder cancer. *Nat. Rev. Urol.*, **10**, 396–404.
25. Oxford, G. and Theodorescu, D. (2003) Review article: the role of ras superfamily proteins in bladder cancer progression. *J. Urol.*, **170**, 1987–1993.
26. Teng, J., Wang, Z.-Y., Jarrard, D.F. and Bjorling, D.E. (2008) Roles of estrogen receptor α and β in modulating urothelial cell proliferation. *Endocr. Relat. Cancer*, **15**, 351–364.
27. Lamm, D.L. (2000) Efficacy and safety of bacille Calmette-Guerin immunotherapy in superficial bladder cancer. *Clin. Infect. Dis.*, **31**, S86–S90.
28. Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. and Stratton, M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
29. Reichardt, L.F. (2006) Neurotrophin-regulated signalling pathways. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **361**, 1545–1564.
30. Carracedo, A., Ma, L., Teruya-Feldstein, J., Rojo, F., Salmena, L., Alimonti, A., Egia, A., Sasaki, A.T., Thomas, G., Kozma, S.C. *et al.* (2008) Inhibition of mTORC1 leads to MAPK pathway activation through a PI3K-dependent feedback loop in human cancer. *J. Clin. Invest.*, **118**, 3065–3074.
31. Santarpia, L., Lippman, S.M. and El-Naggar, A.K. (2012) Targeting the MAPK–RAS–RAF signaling pathway in cancer therapy. *Expert Opin. Ther. Targets*, **16**, 103–119.
32. Davis, J.R. and Morris, R.N. (1963) The isolation of intermediary metabolites of l-lysine-U-C14 from tissues of tumor-bearing rats. *Cancer Res.*, **23**, 143–147.
33. Young, S.Z. and Bordey, A. (2009) GABA's control of stem and cancer cell proliferation in adult neural and peripheral niches. *Physiology*, **24**, 171–185.
34. Menche, J., Sharma, A., Kitsak, M., Ghiassian, S.D., Vidal, M., Loscalzo, J. and Barabási, A.-L. (2015) Uncovering disease-disease relationships through the incomplete interactome. *Science*, **347**, 1257601.
35. Johnston, J.J., Olivos-Glander, I., Killoran, C., Elson, E., Turner, J.T., Peters, K.F., Abbott, M.H., Aughton, D.J., Aylsworth, A.S., Bamshad, M.J. *et al.* (2005) Molecular and clinical analyses of Greig cephalopolysyndactyly and Pallister-Hall syndromes: robust phenotype prediction from the type and position of GLI3 mutations. *Am. J. Hum. Genet.*, **76**, 609–622.
36. Busskamp, V., Duebel, J., Balya, D., Fradot, M., Viney, T.J., Siebert, S., Groner, A.C., Cabuy, E., Forster, V., Seeliger, M. *et al.* (2010) Genetic reactivation of cone photoreceptors restores visual responses in retinitis pigmentosa. *Science*, **329**, 413–417.
37. Barnes, C.S., Alexander, K.R. and Fishman, G.A. (2002) A distinctive form of congenital stationary night blindness with cone ON-pathway dysfunction. *Ophthalmology*, **109**, 575–583.
38. Hosack, D.A., Dennis, G. Jr, Sherman, B.T., Lane, H.C. and Lempicki, R.A. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol.*, **4**, R70.
39. Ogris, C., Helleday, T. and Sonnhammer, E.L.L. (2016) PathwAX: a web server for network crosstalk based pathway annotation. *Nucleic Acids Res.*, **44**, W105–W109.