# Assessment of effectiveness of the network-guided genetic screen

**Eiru Kim, Jaeyoon Shin and Insuk Lee***

*Received 15th April 2010, Accepted 25th June 2010*
**DOI: 10.1039/c005131d**



**Network-guided genetic screen (NGGS) has been suggested to be more effective than traditional forward or reverse genetics screen. Although the predictability of NGGS with given networks has been measured in previous studies, its general effectiveness, particularly for genome-wide reverse genetic screens, has not been assessed yet. We estimated the general effectiveness of NGGS by simulating iterative searching over networks for known phenotypic genes in two model organisms, baker's yeast (*S. cerevisiae*) and worm (*C. elegans*). We found that NGGS is more effective in *C. elegans*, implicating a higher value for NGGS in animals including humans.**

Genetic dissection of traits with medical or agronomic importance is one of the ultimate goals in genetics research. It is however not a trivial task because of the complex genetic organization of organisms. During recent decades, geneticists have exploited two complimentary approaches for genetic analysis of phenotypes: forward and reverse genetic screens. A forward genetic screen is a phenotype-driven approach, in which we typically design a screening system detecting mutants displaying a phenotype of interest. Subsequent genotyping leads to the discovery of novel genes responsible for the phenotype. However, this 'blindfolded' approach suffers from its bias towards strong genetic factors—strong enough to be detected by a given screen method—and genes located on a more sensitive region of the chromosome to the given random mutagenesis method (for example, transposable element mutagens usually show uneven distribution of insertions along the chromosome). Another type of forward genetics is the recently advanced genome-wide association study (for a review, see ref. 1) that shows huge potential in discovery of candidate genes associated with phenotypes of humans and plants. Yet, suggesting only a handful candidate genes with significant statistical power, this new genetics technology does not seem sensitive enough to detect weak genetic factors.[2] On the contrary, an alternative systematic gene-driven reverse genetics approach can potentially circumvent the problem of false negatives by permitting testing of each gene, thereby detecting genes causing even subtle phenotypic effect. In virtue of available genome-wide knock-out mutant libraries (*e.g.*, for yeast and *Arabidopsis*) or gene silencing methods (*e.g.*, RNA interference systems for worm, fly, and mammalian cells), unbiased genome-wide reverse genetic screens are possible. The genome-wide reverse genetic screen, however, is not a pragmatic approach in general, because testing all genes in the genome incurs a high cost as well as frequently missing many genes truly involved in the phenotype due to the high throughput nature of the screens.

Recently a novel strategy for genetic screening with the aid of predictive gene network models has been proposed—network-guided genetic screen (NGGS).[3,4] In NGGS, we prioritize candidate genes for experimental tests by strength of connection to the known phenotypic genes in a gene network—generally called guilt-by-association approach. Many phenotypes are genetically organized as pathways or functional modules that are composed of functionally coupled genes, often forming clusters of highly connected genes in functional gene networks. Therefore, known phenotypic genes tend to connect to novel genes for the same phenotypes in the network. The predictability of a particular phenotype by a network can be measured by cross-validation of the known phenotypic genes with formal Receiver Operating Characteristics (ROC) curve analysis.[4] In this analysis predictability is measured by the area under the ROC curve (AUC) scores spanning from 0.5 indicating predictability by random expectation to 1.0 for a perfect predictor. If a phenotype is predictable by a given network, discovery of novel genes for the same phenotype could be effectively conducted by guilt-by-association. The feasibility of this cost-effective genetic screen has been experimentally validated for phenotypes in various model organisms.[5–7]

Despite those successes of NGGS, its general effectiveness over unbiased genome-wide reverse genetic screens has not been assessed systematically. For example, we achieved 15-fold effectiveness for discovery of ribosomal biogenesis genes of baker's yeast,[7] and 10-fold effectiveness for discovery of suppressors of tumorigenesis pathway mutations in *C. elegans* (worm).[6] We do not know, however, how generally effective the NGGS would be for genome-wide reverse genetic screen projects in yeast and worm. In this study, we measured general cost effectiveness of NGGS over unbiased genome-wide reverse genetic screen for loss-of-function phenotypes of yeast and worm by simulation of iterative process of prediction-and-test. The scheme of computational simulation of an iterative screen based on guilt-by-association followed by a test for known phenotypic genes is illustrated in Fig. 1. For the simulation we used the same gene sets for 100 yeast knock-out phenotypes as in ref. 4 and 43 worm RNA interference (RNAi) phenotypes as in ref. 6. In the simulation we start the screen for genes of each phenotype with a random selection of a candidate. If the random candidate is not a known phenotypic gene (false positive), we go for another random selection. Or if the random candidate is a known phenotypic gene (true positive), we select its network neighbors as the next candidates to test. Subsequently for the candidates that turn out to be true positives, we continue to collect their neighbors as the next novel candidates. On the contrary, for the candidates that

*Department of Biotechnology, College of Life Science and Biotechnology, Yonsei University, 134 Shinchon-dong, Seodaemun-ku, Seoul 120-749, Korea. E-mail: insuklee@yonsei.ac.kr*

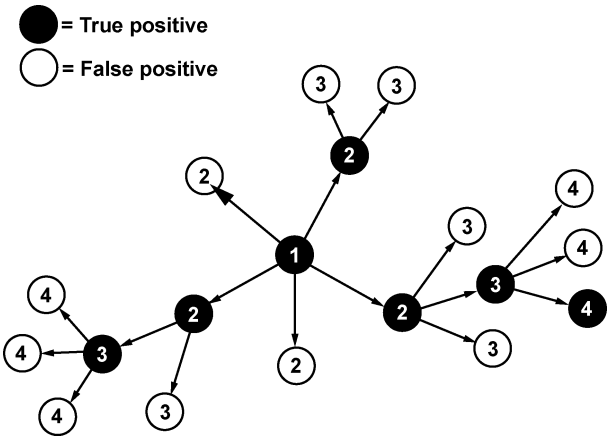


**Fig. 1** A schematic figure describing the iterative steps of prediction-and-test to search for true phenotypic genes in the genome *via* network connections. Numbers inside nodes represent the sequence number of iteration. For example, we choose a #1 node and find it as a true positive (black node). Then we use this node to predict next candidates (all five #2 nodes) by connections to the #1 node. Among five only three turn out to be true positives. We continue to predict the next candidates by network neighbors connected to these true positives, whereas network-guided prediction stops at false positive nodes (white nodes).

turn out to be false positives, we stop propagation of candidates over the network. In case we do not find any phenotypic genes among the candidates, we go back to the initial condition of random selection for next candidates. If phenotypic genes are connected as pathways in a gene network, this rationale of progressive search will effectively extend a view of pathways responsible for the phenotypes (Fig. 1).

Assuming a higher probability of true phenotypic genes for candidates by guilt-by-association, we expect a reduced number of genes to be tested in total to retrieve all known phenotypic genes by NGGS. We formulate a new score to measure effectiveness ($E$) of NGGS over unbiased genome-wide reverse genetic screens as follows.

$$E = \log_2 \frac{R}{N},$$

where $R$ = the number of tested genes to retrieve 50% of total known phenotypic genes *by random selection*, $N$ = the number of tested genes to retrieve 50% of total known phenotypic genes *by NGGS*. Therefore, ineffective NGGS would be indicated by $E$ equal to zero, while effective NGGS by positive $E$ value (for example, $E = 1$ if NGGS requires half as many tested genes for unbiased random selection to retrieve 50% of total known phenotypic genes). Here, we choose 50% retrieval rate for optimal measurement of effectiveness to avoid any lagging effect as approaching saturating point of retrieval.

Using the above scoring scheme, we evaluated the general effectiveness of NGGS in genome-wide reverse genetic screens for loss-of-function phenotypes for which predictability has been previously measured. For a more robust evaluation, we analyzed only phenotypes with not less than 30 associated genes for both yeast and worm (thus, we analyzed 53 knock-out phenotypes of yeast and 28 RNAi phenotypes of worm). As gene network models for NGGS, we used previously reported probabilistic functional networks of yeast genes, YeastNet,[8] and of worm genes, WormNet.[6] The original networks were modified by excluding links supported by co-citation or genetic interaction data that could be directly related to our test data, loss-of-function phenotypic genes. Therefore, we conducted the entire analyses with a highly conservative setting. We found strong correlation between predictability of phenotypic genes (measure by AUC) and effectiveness of NGGS over an unbiased genome-wide reverse genetic screen for the phenotypic genes (measured by $E$) in both yeast and worm (Fig. 2). With a regression model (quadratic fit) between predictability and effectiveness, AUC of 0.7 and 0.85 correspond to 2-fold and 5-fold effectiveness, respectively, in yeast. We observed better correlation with worm in which AUC of 0.67, 0.8, and 0.88 correspond to 2-fold, 5-fold, and 10-fold effectiveness, respectively. For yeast, only 28% (15/53) tested knock-out phenotypes are expected to be 2-fold or more effective by NGGS (Fig. 2a). For worm, however, 68% (19/28) tested RNAi phenotypes are expected to be 2-fold or more effective by NGGS (Fig. 2b). This suggests a generally higher effectiveness of NGGS with a higher eukaryote such as worm that has a larger genome, that is, a larger search space. Maximum effectiveness by maximum predictability (AUC = 1) is about 15-fold for yeast and about 30-fold for worm with the given regression models. These are fairly consistent with the range of effectiveness previously reported from experimental validation in yeast[7] and worm.[6]

A protein–protein interaction network (PPIN) is a major type of pathway model in systems biology. While PPINs depict pathways by physical interactions between proteins, functional gene networks used in this analysis do so by functional association between genes. A physical interaction between proteins is strong evidence of functional association between genes encoding those proteins. Yet, there are many gene functional associations that are not based on physical protein interactions. Moreover, available techniques to detect physical protein interactions suffer from limited sensitivity and high false positive rates, particularly with multicellular organisms such as animals and plants. We compared the effectiveness of NGGS by PPINs and by functional gene networks in yeast and worm. We constructed a PPIN of yeast by consolidating various protein–protein interaction (PPI) data sets: PPI databases such as Database of Interacting Proteins (DIP)[9] (used only small scale experiment set), Munich Information Center for Protein Sequence (MIPS),[10] BioGRID,[11] confident sets of genome-wide high-throughput yeast two hybrid

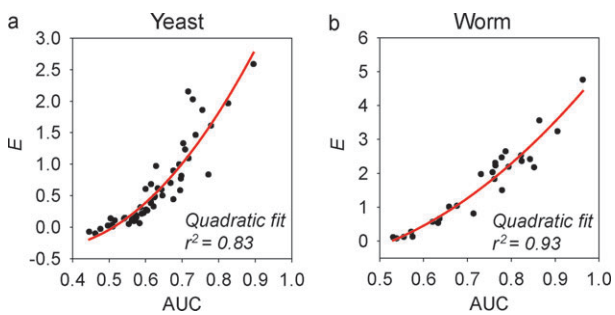


**Fig. 2** Regression models between predictability (AUC) and effectiveness ($E$) of loss-of-function phenotypes measure by (a) YeastNet in yeast and by (b) WormNet in worm.

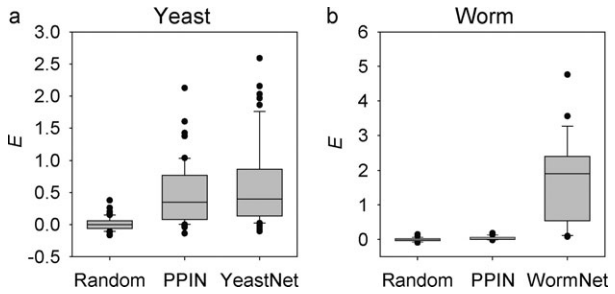


**Fig. 3** Comparison of the general effectiveness by NGGS between functional gene networks (YeastNet or WormNet) and protein–protein interaction networks (PPIN) (a) for yeast and (b) for worm.

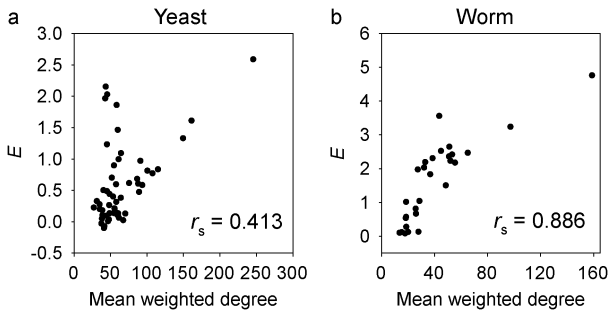


**Fig. 4** Relationship between mean weighted-degree of member genes for each tested loss-of-function phenotype and its effectiveness by NGGS (a) in yeast and (b) in worm. We observed a higher correlation (measured by Spearman rank correlation, $r_s$) between mean weighted-degree and effectiveness of NGGS for worm.



screens,[12,13] and PPIs derived using the spoke model[14] of affinity purification of protein complex followed by mass spectrometry analysis data.[15,16] This PPIN of yeast comprises 5275 genes and 65 033 links. In contrast to yeast, worm has much fewer PPI data. We employed the largest worm PPIN, Worm Interactome version 8 or WI8, based on high throughput yeast two hybrid screens and literature curation.[17] This largest published worm interactome maps only 4391 links between 2812 genes (~14% of genome). The measured effectiveness of NGGS for knock-out phenotypes by PPIN or by functional gene network (YeastNet) shows similar distribution in yeast ($p$-value > 0.99 by Wilcoxon signed rank test) (Fig. 3a). This suggests that the current status of yeast PPIN recapitulates the majority of pathways. On the contrary, with the higher eukaryote worm, the functional gene network (WormNet) is much more effective than the worm PPIN, WI8. As expected from its limited genome coverage, WI8 shows effectiveness of NGGS for RNAi phenotypes that is similar to that of a random model (Fig. 3b). These results suggest that NGGS would be much more powerful with a functional gene network in higher eukaryotes in which PPI data are insufficient for high coverage pathway reconstruction.

What are the characteristics of the loss-of-function phenotypes with high effectiveness of NGGS? In principle, NGGS uses network connectivity among genes associated with same phenotypes. Thus, degree of connectivity of phenotypic genes to the entire network is presumably an important factor promoting effectiveness of NGGS. To address this question, we measured mean degree connectivity of the member genes of each tested loss-of-function phenotype with edge weight of the networks (Fig. 4). We observed a strong tendency for higher effectiveness of NGGS with higher mean degree connectivity of member genes for each phenotype in both yeast and worm. Some yeast phenotypes show high effectiveness even with low degree connectivity (Fig. 4a). This suggests that these phenotypic genes are connected to each other with higher specificity than other phenotypes.

In summary, we assessed the general effectiveness of NGGS over unbiased genome-wide reverse genetic screens by computational simulation in yeast and worm. Predictability strongly correlates with effectiveness of NGGS. For at least 2-fold effectiveness, we need to obtain predictability score, AUC of 0.7 and 0.67 for yeast and worm, respectively. We also estimated maximally achievable effectiveness of 15-fold and 30-fold for yeast and worm, respectively, with currently available gene functional networks. It is noticeable that NGGS is more effective in a larger genome, implicating a higher value for NGGS in animals including human. Further, NGGS can be more effective with functional gene networks than PPINs, especially for organisms with insufficient protein–protein interaction data such as *C. elegans*.

We thank Ben Lehner for critical comments on the manuscript. This work was supported by grants from the National Research Foundation of Korea (NRF) funded by the Korea government (MEST) (No. 2009-0063342, 2009-0070968, 2009-0087951) to I.L.

## References

1 J. N. Hirschhorn and M. J. Daly, *Nat. Rev. Genet.*, 2005, **6**, 95–108.
2 B. Maher, *Nature*, 2008, **456**, 18–21.
3 B. Lehner and I. Lee, *Briefings Funct. Genomics Proteomics*, 2008, **7**, 217–227.
4 K. L. McGary, I. Lee and E. M. Marcotte, *GenomeBiology*, 2007, **8**, R258.
5 I. Lee, B. Ambaru, P. Thakkar, E. M. Marcotte and S. Y. Rhee, *Nat. Biotechnol.*, 2010, **28**, 149–156.
6 I. Lee, B. Lehner, C. Crombie, W. Wong, A. G. Fraser and E. M. Marcotte, *Nat. Genet.*, 2008, **40**, 181–188.
7 Z. Li, I. Lee, E. Moradi, N. J. Hung, A. W. Johnson and E. M. Marcotte, *PLoS Biol.*, 2009, **7**, e1000213.
8 I. Lee, Z. Li and E. M. Marcotte, *PLoS One*, 2007, **2**, e988.
9 I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. M. Kim and D. Eisenberg, *Nucleic Acids Res.*, 2002, **30**, 303–305.
10 H. W. Mewes, S. Dietmann, D. Frishman, R. Gregory, G. Mannhaupt, K. F. Mayer, M. Munsterkotter, A. Ruepp, M. Spannagl, V. Stumpflen and T. Rattei, *Nucleic Acids Res.*, 2008, **36**, D196–201.
11 B. J. Breitkreutz, C. Stark, T. Reguly, L. Boucher, A. Breitkreutz, M. Livstone, R. Oughtred, D. H. Lackner, J. Bahler, V. Wood, K. Dolinski and M. Tyers, *Nucleic Acids Res.*, 2008, **36**, D637–640.
12 T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori and Y. Sakaki, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 4569–4574.
13 P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields and J. M. Rothberg, *Nature*, 2000, **403**, 623–627.
14 G. D. Bader and C. W. Hogue, *Nat. Biotechnol.*, 2002, **20**, 991–997.
15 A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M. A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A. M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell and G. Superti-Furga, *Nature*, 2006, **440**, 631–636.

16 N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrin-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. St Onge, S. Ghanny, M. H. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili and J. F. Greenblatt, *Nature*, 2006, **440**, 637–643.

17 N. Simonis, J. F. Rual, A. R. Carvunis, M. Tasan, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, J. M. Sahalie, K. Venkatesan, F. Gebreab, S. Cevik, N. Klitgord, C. Fan, P. Braun, N. Li, N. Ayivi-Guedehoussou, E. Dann, N. Bertin, D. Szeto, A. Dricot, M. A. Yildirim, C. Lin, A. S. de Smet, H. L. Kao, C. Simon, A. Smolyar, J. S. Ahn, M. Tewari, M. Boxem, S. Milstein, H. Yu, M. Dreze, J. Vandenhaute, K. C. Gunsalus, M. E. Cusick, D. E. Hill, J. Tavernier, F. P. Roth and M. Vidal, *Nat. Methods*, 2009, **6**, 47–54.



 