Identification of Key Causal Regulators in Gene Networks

Bin Zhang and Jun Zhu

Abstract— **One primary goal of gene network analysis is to identify key regulatory components, or key drivers, of sub-networks with respect to various biological contexts. Here we developed a general algorithm to identify key drivers in gene regulatory networks. The generalized key driver analysis (KDA) uncovers not only the well-known regulators for the expression quantitative trait locus (eQTL) hotspots but also many novel drivers underlying certain eQTL hotspots. This algorithm also identifies a number of key regulators for immune response involved in multiple common chronic diseases and the predicted drivers appear to be more important than the non-drivers genes to induce phenotypic changes when perturbed.**

Index Terms— **causal networks, gene regulatory networks, key drivers**

I. Introduction

Inferring causal-reactive relationships between genes has been an important topic of systems biology. With the increasing availability of large scale genomic and genetic data, many gene regulatory networks have been reconstructed and tested [1-5]. A common practice for testing the prediction power of such networks involves the identification of key regulators for groups of functionally relevant genes, followed by biological validation of the effect of perturbing the putative regulators [6].

In a previous work of dissecting expression quantitative trait loci (eQTL) hot spots using Bayesian networks [6], a procedure called key driver analysis (KDA) was used to infer the causal regulators for these hot spots. The procedure is briefly described: (1) for each eQTL hotspot region, *cis* eQTLs are selected as putative regulators; (2) downstream genes from the putative regulators are selected from the Bayesian network and intersected with the eQTL hotspot genes; (3) a statistically significant overlap identifies the putative regulator as a key driver of the hot spot. When applied to a yeast regulatory network, KDA uncovered all previous known regulators within 8 of the 13 eQTL hotspots as well as new regulators, which were validated experimentally. However, there are several shortcomings with this implementation: (i) it limits to pre-selected candidate drivers (e.g. cis-eQTL genes), (ii) considering all the downstream nodes of a candidate driver for an enrichment test may not be optimal, and (iii) for complex networks searching the whole network is not necessary and computationally expensive. To resolve these problems with the original approach and to make it more broadly useful, here we formally defined the KDA algorithm.

II. Methods and Materials

Key driver analysis (KDA) takes as input a set of genes (**G**) and a directed gene network N. The objective is to identify the key regulators for the gene sets with respect to the given network. Figure 1 shows the general procedure of KDA. KDA first generates a sub-network N_G, defined as the set of nodes in N that are no more than h-layers away from the nodes in **G**. Two extreme cases are: (i) h=0, i.e., N_G consists of only the links among the nodes from G; (ii) N_G = N, i.e., take the whole network as N_G. For a dense network N, we recommend case (i) in order to derive a simpler subnetwork for the subsequent analysis. Depending on whether the set of nodes in N_G is a subset of G or not, we use either a static or dynamic neighborhood search. The dynamic neighborhood search (DNS) searches the h-layer neighborhood (h=1,..,H) for each gene in N_G ($HLN_{g,h}$) for the optimal h*, such that

$$ES_{h*} = \max(ES_{h,g})\,\forall g \in N_g, h \in \{1..H\}$$

where $ES_{h,g}$ is the computed enrichment statistic for $HLN_{g,h}$. The static neighborhood search (SNS), on the other hand, considers only a pre-specified h-layer neighborhood.

In DNS, a node becomes a candidate driver if its HLN is significantly enriched for the nodes in G. Note that here the enrichment test is computed using the subnetwork N_G as background. Candidate drivers without any parent node (i.e., root nodes in directed networks) are designated as global drivers and the rest are local drivers. We also promote as global drivers the nodes whose HLN is most significantly enriched for the signature by taking the whole network N as the background; this process is called HLN outlier detection. Specifically, we first test HLNs for the enrichment of the signature against the whole network. As the enrichment tests against different backgrounds don’t always agree with each other, this step basically rescues those master nodes missed by enrichment tests against a subnetwork.

In SNS, candidate drivers are identified as follows. We first compute the size of the h-layer neighborhood (HLN) for each node. For the given network N, let $\boldsymbol{\mu}$ be the sizes of HLNs and $\boldsymbol{d}$ be the out-degrees for all the nodes. The nodes with the sizes of their HLN greater than $\overline{\mu} + \sigma(\mu)$ are nominated as candidate drivers. The candidate drivers

Manuscript received March 06, 2013; revised April 7, 2013.

B. Zhang (corresponding author) is with the Department of Genetics and Genomic Sciences and the Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA (phone: 212-824-8947; fax: 646-537-8660; e-mail: bin.zhang@mssm.edu; web: www.mssm.edu/profiles/bin-zhang).

J. Zhu is with the Department of Genetics and Genomic Sciences and the Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA (Email: jun.zhu@mssm.edu).

TABLE I

KEY DRIVERS IDENTIFIED BY THE ORIGINAL AND THE GENERALIZED KDA APPROACHES. FOR THE GENERALIZED KDA, WE SHOW RESULTS FROM DIFFERENT LAYER NEIGHBORHOODS (L1: 1 LAYER, L2: 2 LAYERS, L3: 3 LAYERS).

eQTL hotspot	Hotspot chr.	Hotspot base-pair position	the original KDA (Zhu, Zhang et al. 2008)	KDA L1	KDA L2	KDA L3
2	2	560000	TBS1, TOS1, ARA1, CSH1, SUP45, CNS1, AMN1	TBS1, ARA1, CSH1, SUP45, CNS1, PWP2	TBS1, TOS1, ARA1, CSH1, SUP45, CNS1, ENP2, NOP7	TBS1, TOS1, ARA1, CSH1, SUP45, CNS1, NMD3, RPF1
4	3	1.00E+05	LEU2, ILV6, NFS1, CIT2, MATALPHA1	LEU2, BAP2, OAC1	BAP2, LEU2, OAC1, RTG3	LEU2, BAP2, OAC1, RTG3
5	3	230000	MATALPHA1			
6	5	130000	URA3	URA3	URA3	URA3
7	8	130000	GPA1	GPA1	GPA1	GPA1
8	12	680000	HAP1	HAP1	HAP1	HAP1
9	12	107000	YRF1-4, YRF1-5, YLR464W	YRF1-4	YRF1-4	YRF1-4
11	14	503000	SAL1, TOP2	SAL1, RSM24, RSM25	SAL1, RSM24, RSM25, MRPL3	SAL1, RSM24, RSM25, MRPL3
12	15	180000	PHM7	TFS1, PHM7, TKL2, YGR052W	PHM7, TFS1, YGR043C, HXT7, TKL2, GDB1, YGR052W	PHM7, TFS1, YGR043C, PIL1, TKL2, HXT7, YGR052W, GDB1
10	13	70000		GCV1	GCV1	GCV1
13	15	590000		ATP5	ATP20	ATP5, ATP20

without any parent node (i.e., root nodes) are nominated as global drivers. Similar to DNS, we also promote hub nodes as global drivers, i.e., the nodes with out-degrees above $\overline{d} + 2\sigma(d)$ are designated as global drivers.

III. RESULTS

We evaluate the performance of KDA on a yeast and an inflammation disease studies. In both studies predictive Bayesian networks were constructed. We applied KDA to the yeast causal network to predict causal regulators responsible for hot spots of gene expression activity in a segregating yeast population, and to human and mouse tissue networks to identify key regulators of immune response associated with common chronic diseases.

A. *Key drivers of eQTL Hot spots in Yeast*

In this application, we considered a genotypic and expression data from a yeast cross of 112 segregants constructed from the BY and RM strains of *S. cerevisiae* (referred to here as the BXR cross) [7]. A previous genome-wide genetic linkage analysis mapped expression quantitative trait loci (eQTL) for each of the 5,740 expression traits represented on the microarray and identified 13 chromosomal regions harboring a large number of eQTL, i.e., eQTL hot spots [6]. While many studies have been conducted on this particular dataset to predict the drivers of the eQTL hot spots by inferring causal relationship between genes under the control of specific genetic loci [4-6, 8], a Bayesian network reconstructed by integrating genotypic, gene expression, protein-protein interaction and transcription factor binding site (TFBS) data remains the most predictive [6]. Therefore, we applied KDA on the most predictive Bayesian network.

As the yeast Bayesian network is quite sparse (the average number of links per node is 2.2), KDA was based on DNS. We compared the results by KDA and the original implementation and tested how robust the results are with respect to the expansion (different layer neighborhoods (L1-1 layer, L2-2 layers, L3-3 layers). As shown in Table 1, all the major regulators predicted by the original approach were also uncovered by KDA-L1, KDA-L2 and KDA-L3 except a few very weak regulators (whose neighborhoods are not highly enriched the genes linking to the corresponding eQTL hot spots) such AMN1, MATALPHA1 and TOP2. Notably, KDA uncovered many new putative trans-QTL regulators. For the hot spot 11, KDA identified three new regulators, RSM24, RSM25 and MRPL3 in addition to SAL1 which is the only one predicted by the original approach [6]. Figure 2 shows the subnetwork and the key drivers for the hot spot 11. 277 genes link to the hot spot 11

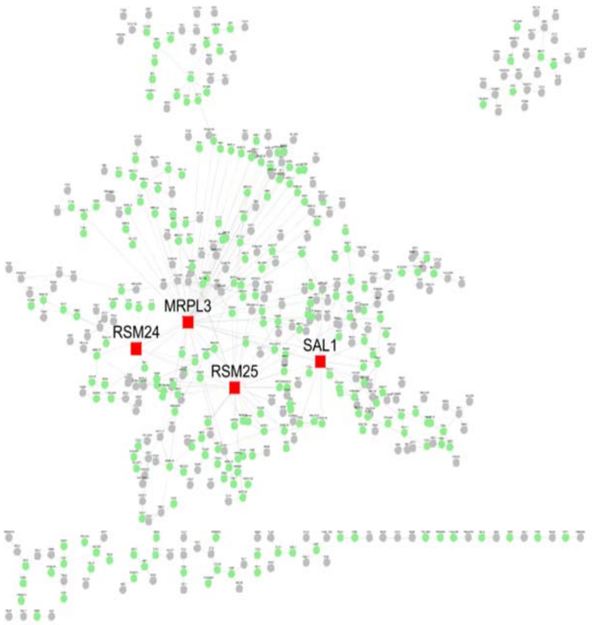

Fig. 1. A Bayesian subnetwork and the key drivers for the eQTL hotspot 11 (Chr14, 503000bp) in a yeast cross. The large square nodes (SAL1, RSM24, RSM25, MRPL3) are the global drivers.

and 242 of them were included in the subnetwork based on a 3-layer expansion. While only 98 of the 242 genes are the downstream of SAL1 (p<7e-95), 142 are the downstream of RSM25 (p<1.42e-114). Figure 1 shows these four key drivers in the subnetwork. For the hot spot 12, KDA uncovered 6 new regulators in addition to PHM7 which is the only regulator identified by the original approach. Moreover, the neighborhoods of TFS1, YGR043C, TKL2 and YGR052W are more significantly enriched for the genes links to the hot spot than that of PHM7. For example, 63% (83) of the 132 3-LN nodes of TFS1 have eQTL on the hot spot 12 with which 340 genes are associated (p<3.5e-54) while 40 of the 46 2-LN nodes of PHM7 are linked to the same hot spot (p<2e-35). For the hotspots 10 and 13, no regulator was identified by the original KDA but the generalized KDA predicted GCV1 as a regulator for the hot spot 12, and ATP5 and ATP20 as regulators for the hot spot 13. 5 of the 6 2-LN nodes of GCV1 have eQTL on the hotspot 10 which includes 41 eQTL genes (p<8.3e-10). 16 of the 35 6-LN nodes of ATP20 have eQTL on the hotspot 13 which includes 33 eQTL genes (p<4.3e-26). Table 1 also shows the robustness of the KDA with respect to the selection of the parameter of the expansion range (layer).

B. *Key drivers of Immune Response in Chronic Inflammation Diseases*

Previously, we built up a set of tissue-specific consensus Bayesian networks[9] from several large scale genetic and genomic studies of complex diseases [10-13]. These consensus BNs and a common inflammatome gene signature identified from multiple inflammatory disease models were then used as input for KDA to identify 151 key drivers for inflammation response[9]. Figure 2 shows an inflammation regulatory network conserved in both human adipose and human liver and the predicted key drivers are highlighted.

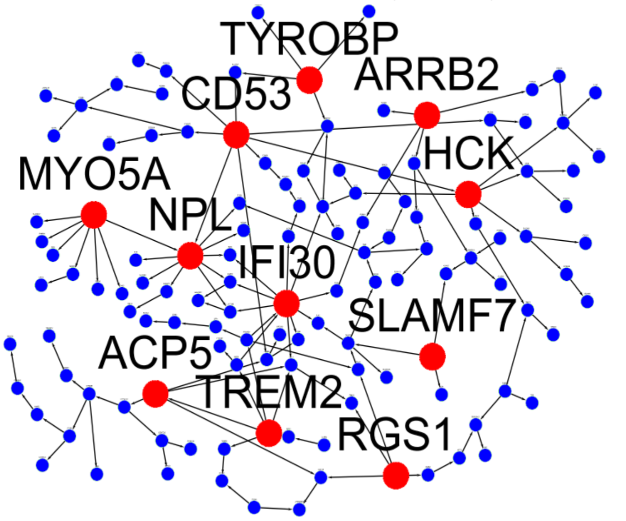

Fig. 2. The key drivers and a consensus Bayesian subnetwork for immune response. The key drivers in the network are highlighted in larger size.

We utilized the mutant phenotype data from the Mouse Genome Informatics database (MGI) to validate the predicted key drivers. In the MGI database, 28.7% of the tested genes give rise to observable altered phenotypes when perturbed. Strikingly, 63.6% of the predicted key drivers have mutant phenotypes (Fisher Exact Test p = $2e^{-12}$). However, only 39.2% of the non-driver genes have mutant phenotypes. Notably, 19 of the top 55 key drivers were tested in MGI and 73.7% (14) had mutant phenotypes (Figure 3). Thus, the key drivers identified through the proposed key driver analysis indeed appear to be more biologically important than the non-drivers.

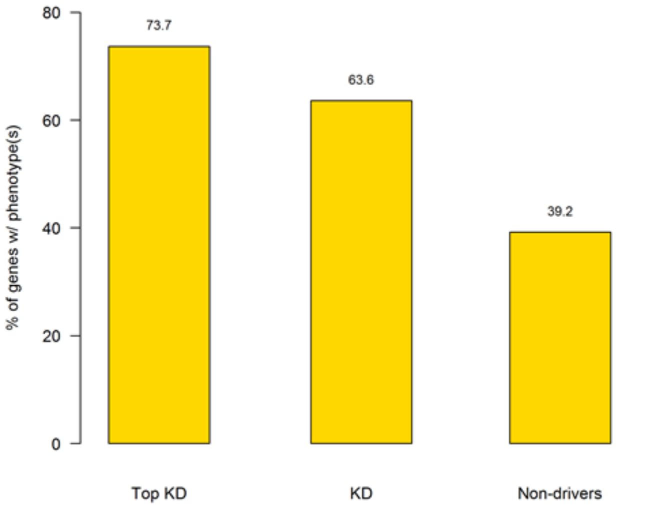

Fig. 3. Percentage of the genes with mutant phenotypes in different gene sets from the key driver analysis on the inflammatome signature and a consensus Bayesian network. An unambiguous trend is that the key drivers are more likely to have mutant phenotypes than the non-driver genes. In particular, the top 55 key drivers have the largest frequency (73.7% as oppose to that (39.2) of the non-drivers.

IV. CONCLUSION

We developed a general key driver analysis algorithm to identify key regulators for a particular gene set of interest with respect to a given regulatory network. To deal with the complexity of gene regulatory networks, the algorithm incorporated a couple of mechanisms such dynamic and static neighborhood search and combination of distinct network connectivity measurements etc. The generalized KDA uncovers not only the well-known regulators for the expression quantitative trait locus (eQTL) hotspots but also many novel drivers underlying certain eQTL hotspots. We also applied this algorithm to uncover a number of regulators for immune response involved in multiple common chronic diseases and the predicted drivers appear to be more biologically important than the non-drivers genes.

REFERENCES

[1]. 1. Schadt, E.E., et al., *An integrative genomics approach to infer causal associations between gene expression and disease.* Nat Genet, 2005. 37(7): p. 710-7.

[2]. 2. Zhu, J., et al., *An integrative genomics approach to the reconstruction of gene networks in segregating populations.* Cytogenet Genome Res, 2004. 105(2-4): p. 363-74.

[3]. 3. Zhang, B., et al., *A trust region method in adaptive finite element framework for bioluminescence tomography.* Opt Express, 2010. 18(7): p. 6477-91.

[4]. 4. Bing, N. and I. Hoeschele, *Genetical genomics analysis of a yeast segregant population for transcription network inference.* Genetics, 2005. 170(2): p. 533-42.

[5]. 5. Kulp, D.C. and M. Jagalur, *Causal inference of regulator-target pairs by gene mapping of expression phenotypes.* BMC Genomics, 2006. 7: p. 125.

[6]. 6. Zhu, J., et al., *Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks.* Nat Genet, 2008. 40(7): p. 854-61.

[7]. 7. Yvert, G., et al., *Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors.* Nat Genet, 2003. 35(1): p. 57-64.

[8]. 8. Millstein, J., et al., *Disentangling molecular relationships with a causal inference test.* BMC Genet, 2009. 10(1): p. 23.

[9]. 9. Wang, I.M., et al., *Systems analysis of eleven rodent disease models reveals an inflammatome signature and key drivers.* Mol Syst Biol, 2012. 8: p. 594.

[10]. 10. Chen, Y., et al., *Variations in DNA elucidate molecular networks that cause disease.* Nature, 2008. 452(7186): p. 429-35.

[11]. 11. Lum, P.Y., et al., *Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes.* J Neurochem, 2006. 97 Suppl 1: p. 50-62.

[12]. 12. Emilsson, V., et al., *Genetics of gene expression and its effect on disease.* Nature, 2008. 452(7186): p. 423-8.

[13]. 13. Yang, X., et al., *Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver.* Genome Res, 2010. 20(8): p. 1020-36.