BIOINFORMATICS DISCOVERY NOTE

Vol. 30 no. 3 2014, pages 305–309
doi:10.1093/bioinformatics/btt683

Systems biology

Advance Access publication November 20, 2013

Identification of transcription factors for drug-associated gene modules and biomedical implications

Min Xiong, Bin Li, Qiang Zhu, Yun-Xing Wang and Hong-Yu Zhang*

National Key Laboratory of Crop Genetic Improvement, Center for Bioinformatics, College of Life Science and Technology, Huazhong Agricultural University, Wuhan 430070, P. R. China

Associate Editor: Martin Bishop

ABSTRACT

Motivation: One of the major findings in systems biomedicine is that both pathogenesis of diseases and drug mode of action have a module basis. However, the transcription factors (TFs) regulating the modules remain largely unknown.

Results: In this study, by using biclustering approach FABIA (factor analysis for bicluster acquisition), we generate 49 modules for gene expression profiles on 1309 agent treatments. These modules are of biological relevance in terms of functional enrichment, drug–drug interactions and 3D proximity in chromatins. By using the information of drug targets (some of which are TFs) and biological regulation, the links between 28 modules and 12 specific TFs, such as estrogen receptors (ERs), nuclear factor-like 2 and peroxisome proliferator-activated receptor gamma, can be established. Some of the links are supported by 3D transcriptional regulation data [derived from ChIA-PET (chromatin interaction analysis using paired-end tags) experiments] and drug mode of action as well. The relationships between modules and TFs provide new clues to interpreting biological regulation mechanisms, in particular, the lipid metabolism regulation by ERα. In addition, the links between natural products (e.g. polyphenols) and their associated modules and TFs are helpful to elucidate their polypharmacological effects in terms of activating specific TFs, such as ERs, nuclear factor-like 2 and peroxisome proliferator-activated receptor gamma.

Contact: zhy630@mail.hzau.edu.cn

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on September 5, 2013; revised on November 3, 2013; accepted on November 18, 2013

1 INTRODUCTION

Roughly speaking, the basis of drug actions is to rectify the gene expression disorders in diseases through modulating transcription factors (TFs). Therefore, the gene expression-based analysis has played a critical role in drug repurposing, lead discovery and elucidating drug mode of action (MoA) (Iorio *et al*., 2010). The pathogenesis of diseases has been revealed to be tightly connected to gene modules, which are sets of functionally related genes (Suthram *et al*., 2010). Intriguingly, by using biclustering approach iterative signature algorithm, Iskar *et al.* revealed that drug MoA also has a module basis, which provided new insights into drug actions and perturbed cellular systems as well (Iskar *et al*., 2013). However, the TFs regulating the drug-associated transcriptional modules remain unknown, which prevents us to understand the molecular mechanisms underlying the drug-induced transcriptional variations. Considering the fact that some drug targets are TFs, such as estrogen receptors (ERs), peroxisome proliferator-activated receptors (PPARs) and glucocorticoid receptor (GR), we attempt to use the target information of the module-coupled drugs to identify the TFs. Because biclustering approach FABIA (factor analysis for bicluster acquisition) has shown good performance in recognizing biclusters and has been successfully used in identifying gene expression modules and associated TFs (Gu and Liu, 2008; Hochreiter *et al*., 2010), we used FABIA to generate drug-induced transcriptional modules. In this study, a module is not only a set of genes but also connected with a set of agents. So, a module is exactly a bicluster. Then, we demonstrate that the modules are of biological relevance in terms of functional enrichment, drug–drug interactions (DDIs) and 3D proximity in chromatins. Finally, we identify the TFs for some modules, by which we reveal some new mechanisms for metabolic regulation, and elucidate the polypharmacological mechanisms for some natural products.

2 METHODS

2.1 Data preprocessing

The raw data were downloaded from connectivity map (cMap) (Lamb *et al*., 2006), which consist of 7056 gene chips corresponding to five cultured human cell lines treated with 1309 agents.

Each chemical treatment chip had a vehicle control pair. The data were first normalized by Robust Multi-array Average expression measure. Amplitude (a), a parameter merging the treatment and control data, was used to measure the extent of the differential expression of a given probe set (Lamb *et al*., 2006). The definition of a is as follows:

$$a = \frac{t-c}{(t+c)/2} \tag{1}$$

where t is the expression value for the treatment and c is the expression value for the control. If $a > 0$, the expression is increased on treatment; if $a < 0$, the expression is decreased on treatment. However, if $a = 0$, no change in the expression is observed.

For each agent, the median of expression values determined under different conditions was applied to represent the expression profile. This merged the 7056 expression profiles into 1309 agent-specific transcriptional response profiles, constituting a matrix of 22 215 rows (probes)

*To whom correspondence should be addressed.

and 1309 columns (agents). Each column was normalized using the following equation:

$$x^{*}_{ij} = \frac{x_{ij} - x_j}{s_j} \tag{2}$$

where x_{ij} is the expression value in row i and column j, x_j is the mean value of column j and s_j is the standard deviation of column j.

2.2 Biclustering analysis

The FABIA method was used to perform biclustering analysis (Hochreiter *et al.*, 2010). The outer product λz^{T} of two sparse vectors with p bicluster results in a matrix X. The model for the matrix X is as follows:

$$X = \sum_{i=1}^{P} \lambda_i z_i^{\mathrm{T}} + \gamma \tag{3}$$

where the real numbers λ_i, z_i^T and γ correspond to a prototype column vector for genes, sample participating in the i bicluster and the additive noise, respectively. FABIA 2.2.2 software was used to search K biclusters of 22 215 × 1309 matrix. K (number of biclusters) was set to 50. The sparseness factor was set to 0.1 and the number of iterations was set to 10 000. When $\mathrm{K} \geq 49$, the biclusters covered all the information contents of the matrix. To simplify the presentation of bicluster results, we used 1 or 0 to represent the coupling or not of a bicluster to a specific agent, where 1 implies the z values > 0.

2.3 Gene function enrichment

Database for Annotation, Visualization and Integrated Discovery (Huang *et al.*, 2009) was used to indicate Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) information associated with genes in modules, respectively.

2.4 Drug target enrichment for modules

The drug target enrichment for each module was evaluated by a cumulative hypergeometric test (Iorio *et al.*, 2010). The *P*-value was calculated using the following equation:

$$P(x \geq j) = \sum_{i=j}^{\infty} \frac{\binom{K}{i}\binom{M-K}{N-i}}{\binom{M}{N}} \tag{4}$$

where N is the total number of drugs used in the target enrichment significance evaluation (i.e. 477), M is the number of drugs in the module, i is the number of drugs sharing the same target in N and K is the number of drugs sharing the same target in M. Thus, we can calculate the probability by chance, at least x occurrences of a target among those associated with the module.

3 RESULTS AND DISCUSSIONS

We used the cMap as the starting database, which contains 7056 genome-wide expression profiles of five different human cell lines treated with 1309 chemical agents at different dosages (Lamb *et al.*, 2006). Because most of the drug-induced transcriptional modules are conserved across cell lines (Iskar *et al.*, 2013), the 7056 expression profiles were merged into 1309 agent-specific profiles. Then, the FABIA software (Hochreiter *et al.*, 2010) was used to search K biclusters that correspond to the functional modules. When K was ≥49, superfluous biclusters were pushed toward zero, which means that the modules contained all the information from the 1309 samples (Supplementary Fig. S1). Thus, the present analysis was based on 49 modules, which consisted of 5921 probes and were ordered according to their information contents. That is, module 1 contains the richest information and module 49 contains the poorest. On average, each module covers 324 ± 10.6 agents, each agent is associated with 12.1 ± 5.3 modules (Supplementary Tables S1 and S2). Because different modules have different biological functions (see later in the text for details), the module profiles of the agents allow us to dissect their polypharmacological mechanisms.

The biological functions of each module were defined by enriched genes recorded in GO and KEGG pathways. The 49 modules have significant GO enrichments, with 46 having significant KEGG enrichments (Supplementary Table S1). Most modules concentrate on a few functions, and both GO terms and KEGG pathway annotations match well with each other, indicating the functional convergence of the modules.

Further evidence to demonstrate the biological relevance of gene modules comes from DDIs. The occurrence of DDIs means that the pharmacologic effect of a given drug is influenced by the action of another. Drug pairs that form DDIs usually have a high probability to share common targets (Huang *et al.*, 2013). Through searching DrugBank (Wishart, 2008), we obtained 7516 pairs of DDIs. It is interesting to note that DDIs are significantly enriched for drugs covered by the same module (Supplementary Table S3, cumulative hypergeometric test), indicating the strong pharmacologic associations among the drugs.

To investigate whether the modules are associated with transcriptional regulation, we used the data of interaction regions for RNA polymerase II (RNAPII) determined by chromatin interaction analysis using paired-end tags (ChIA-PET) (Li *et al.*, 2012), which defined close and distant regulatory elements directly interacting with gene promoters. The raw data were downloaded from Encyclopedia of DNA Elements ChIA-PET datasets (Li *et al.*, 2012), which consist of data for five different human cell lines and were merged for use in this study. The genes of each module were mapped into the anchor regions of RNAPII (i.e. 2.5 kb upstream and downstream of interaction regions for RNAPII), according to the procedure reported previously (Sandhu *et al.*, 2012). Promoter–promoter interactions were significantly enriched for the genes within the same module (Supplementary Fig. S2, cumulative hypergeometric test), implying that these genes tend to be proximate in 3D structure of chromatins and thus to be transcribed cooperatively (Sandhu *et al.*, 2012).

The target information of module-coupled drugs was used to identify the TFs linked to the modules. We retrieved 573 approved drugs (hitting 536 targets) from 1309 agents by searching DrugBank (Wishart, 2008) and Therapeutic Target Database (Zhu *et al.*, 2012) and found that 209 targets were shared by at least two drugs. These targets and corresponding 477 drugs can be used for target enrichment to validate the coupling between targets and gene modules (Iorio *et al.*, 2010). The results show that 168 targets can be enriched for 49 modules ($P < 0.01$, $q < 0.01$) (Supplementary Table S4, cumulative hypergeometric test), in which 11 are TFs, corresponding to 28 modules (Fig. 1). In particular, the links between ERα and its associated modules can be further validated by the interactions between ERα and its target promoters.

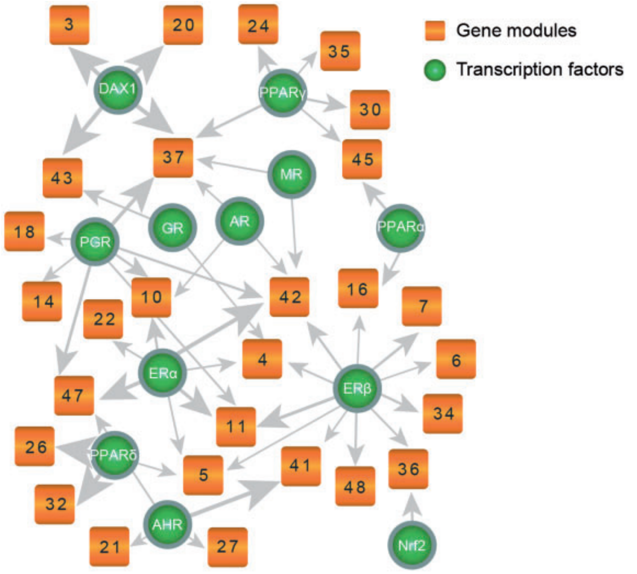

Fig. 1. Transcription factors (TFs, in green) linked to drug-induced transcriptional modules (in orange). By using the target information of module-coupled drugs and by retrieving information of biological regulation, 12 TFs were identified for 28 modules. The size of the arrow represents the negative logarithm of *P*-value in cumulative hypergeometric test. AHR: aryl hydrocarbon receptor; AR: androgen receptor; DAX1: dosage-sensitive sex reversal-adrenal hypoplasia congenita (AHC) critical region on the X chromosome gene 1; ERα/β: estrogen receptor alpha/beta; GR: glucocorticoid receptor; MR: mineralocorticoid receptor; Nrf2, nuclear factor (erythroid-derived 2)-like 2; PGR: progesterone receptor; PPAR$\alpha/\gamma/\delta$: peroxisome proliferator-activated receptor alpha/gamma/delta

It has been broadly accepted that although genomic information is recorded in a linear series of bases, gene regulation occurs in 3D structure of chromatins (Fullwood *et al*., 2009). Because ChIA-PET can efficiently determine the spatial enhancer–promoter interactions, the ChIA-PET-derived long-range chromatin interactions between the ERα-bound regions and their target promoters are appropriate to verify the ERα module relationships (Fullwood *et al*., 2009). We mapped the genes in 49 modules into the anchor regions bound by ERα [i.e. 20 kb within transcription start site to interaction regions, defined by Fullwood *et al*. (2009)] and found that module 42 is richest in potential target genes regulated by ERα, with 66.4% (99/149) probes being potential ERα targets (Supplementary Fig. S2). Modules 22 and 47 are also rich in ERα-regulated genes ($P < 0.00001$, cumulative hypergeometric test) (Supplementary Fig. S2). These three modules are assigned to ERα by drug target enrichment.

It is interesting to note that 7 ERα agonist drugs contained in 1309 agents are all coupled with module 42 (Supplementary Table S5), which provides further evidence to support the link between module 42 and ERα regulation. Thus, it is reasonable to infer that other agents covered by module 42 are highly possible to stimulate ERα. Through searching the adverse drug events database MetaADED (http://www.lmmd.org/online_services/metaadedb/), it was found that 22 of 93 approved drugs covered by module 42 have side effect of gynecomastia (Supplementary Table S6), supporting their ERα stimulation potential. Besides, the agents enriched by module 42 include 9 plant polyphenols (i.e. butein, (−)−catechin, genistein, hesperetin, kaempferol, luteolin, naringenin, nordihydroguaiaretic acid and resveratrol) (Supplementary Table S5), in good agreement with their widely recognized biological effects as phytoestrogens (Kuiper *et al*., 1998).

As illustrated in Figure 1, some modules are coupled with multiple TFs. For instance, module 4 is regulated by ERα, ERβ and GR. GR is well known for its regulation role in lipid metabolism (Yu *et al*., 2010). Module 4 is tightly linked to lipid metabolism, with 37 of 65 genes involved in this process, according to KEGG and GO annotations ($P < 0.0001$) (Supplementary Table S1). Thus, it is reasonable to infer that ERα also plays a role in lipid metabolism regulation. In all, 5 of the 37 genes (i.e. *ACLY*, *FASN*, *ELOVL6*, *MVD* and *DHCR7*) have been reported to participate in lipid metabolism and be regulated by ERα (Villa *et al*., 2012). The remaining 32 genes are thus tentatively inferred to be regulated by ERα in a direct or indirect manner. The ChIA-PET data for ERα indicate that 3 of 32 genes, i.e. *AACS*, *NPC1* and *FDFT1*, are directly regulated by ERα. Furthermore, we established the protein–protein interaction relationships of 37 genes by extracting the protein–protein interaction information (including known and predicted) from the STRING network with intermediate to high confidence (the default STRING evidence scores > 0.4) (Franceschini *et al*., 2013). As shown in Figure 2, there exist strong connections between the five recognized genes and other genes involved in lipid metabolism, supportive of the ERα regulation of these genes. Thus, the TF-based analysis of gene responses to drugs is helpful to reveal new mechanisms for metabolic regulation and produce testable hypotheses.

For TFs that have not been used as drug targets, we can establish their links with modules by retrieving information of biological regulation. According to KEGG annotations, module 36 is tightly associated with glutathione metabolism ($P < 0.0001$) (Supplementary Table S1), which is regulated by nuclear factor (erythroid-derived 2)-like 2 (Nrf2). This module contains the genes for glutathione peroxidase 2 (*GPX2*) and ferritin heavy polypeptide 1 (*FTH1*), as well as all of the antioxidant genes controlled by Nrf2, including *GCLM*, *GCLC*, *NQO1*, *HMOX1*, *GSR* and *PRDX1* (Kaidery *et al*., 2013). Thus, there exists a close link between module 36 and Nrf2, which means that the 189 agents coupled with this module are potential redox regulators (Supplementary Table S7). Some well-known antioxidants, such as ascorbic acid and ebselen, are included in this list. Moreover, the 189 agents involve 4 para-quinones (i.e. tanespimycin, 1,4-chrysenequinone, menadione and tetroquinone), which abstract electrons and protons from the thiols of Keap1, an important negative regulator of Nrf2, resulting in the activation of Nrf2 (Abiko *et al*., 2011).

Six polyphenols, i.e. quercetin, butein, nordihydroguaiaretic acid, myricetin, kaempferol and genistein, are included in this list, suggesting that these polyphenols can serve as redox regulators through activating Nrf2 in relatively low concentrations (low micromolar, recorded in cMap). Although natural polyphenols have long been touted as direct radical scavengers, accumulating evidence indicates that their health benefits are not likely to come from radical-neutralizing activity because their poor

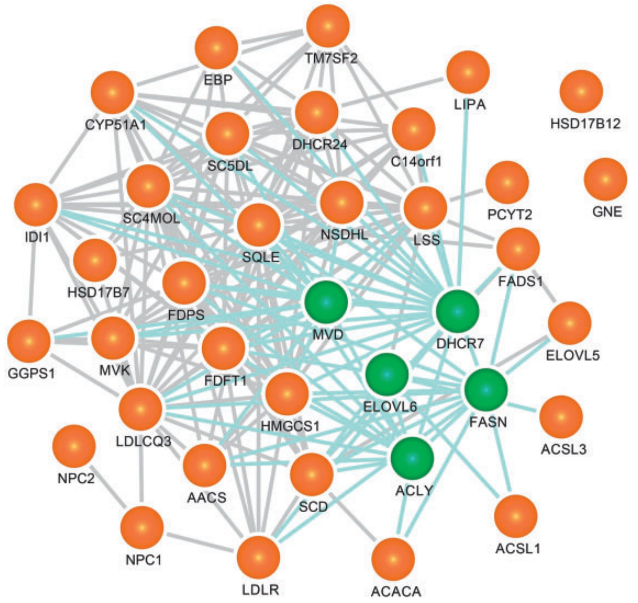

Fig. 2. Protein–protein interaction network of 37 lipid metabolism-associated genes in module 4 (derived from STRING). Five of the genes (in green) have been reported to participate in lipid metabolism and be regulated by ERα (Villa *et al.*, 2012). Others (in orange) are also likely to be regulated by ERα in a direct or indirect manner

bioavailability prevents them to reach high concentrations (high micromolar) to compete with *in vivo* antioxidants such as vitamins C and E (Halliwell *et al.*, 2005). Therefore, the present finding provides solid evidence to support the opinion that natural polyphenols are more likely indirect antioxidants than direct radical modulators (Halliwell *et al.*, 2005; Zhang *et al.*, 2010).

Interestingly, genistein, nordihydroguaiaretic acid, butein and kaempferol are also enriched by module 42, which means that these polyphenols have both antioxidant and estrogen effects. Previous studies have revealed that the side effects of ERα agonists are associated with the stimulated production of reactive oxygen species (Okoh *et al.*, 2013). Thus, antioxidant and estrogen-like activities of natural polyphenols may be responsible for their safety in long-term consumption. Besides modules 42 and 36, these polyphenols are linked to many other modules (Supplementary Table S2), in which modules 11, 48, 10, 25, 3 and 37 are most common and are linked to ERs, DAX1, PPARγ and other TFs (Fig. 1). Although it remains unknown whether polyphenols activate DAX1, genestein and kaempferol have been recognized as PPARγ agonists (O'Leary *et al.*, 2004; Zhang *et al.*, 2013), and thus are cell cycle and apoptosis modulators with chemoprevention potential (Theocharis *et al.*, 2004). Therefore, the well-known health benefits of polyphenols including estrogen-like, antihyperlipidemia and anticancer effects (Gil-Izquierdo *et al.*, 2012; Messina *et al.*, 2009; Tham *et al.*, 1998) can be explained, at least in part, in terms of their potentials in activating specific TFs, such as ERs, Nrf2 and PPARγ.

In summary, by using drug target information and accumulated biological regulation information, we can establish the links between drug-associated gene modules and some TFs. Some of the links are supported by 3D transcriptional regulation data (derived from ChIA-PET experiments) and drug MoA as well. Based on the complex relationships between modules and TFs, we can get new insights into biological regulation, in particular, the lipid metabolism regulation by ERα. Because natural agents usually stimulate multiple TFs, the polypharmacological mechanisms for these agents (e.g. polyphenols) can be partially elucidated in terms of activation of specific TFs. However, it should bear in mind that the present analysis was performed by merging expression profiles of different cell lines, which means that the present results revealed the fundamental links between drugs, TFs and regulated genes, rather than the cell line-dependent features of drug MoA. A more detailed exploration on this topic will be accomplished with the accumulation of biological regulation information, which will offer deeper insights into drug MoA.

ACKNOWLEDGEMENTS

M.X. performed research, analyzed data and wrote the article. B.L. performed research and analyzed data; Q.Z. performed research; Y.X.W. performed research; H.Y.Z. designed research, analyzed data and wrote the article.

Funding: National Basic Research Program of China (973 project, grant 2010CB126100), the National Natural Science Foundation of China (grant 21173092) and the Natural Science Foundation of Hubei Province (grant 2013CFA016).

Conflict of Interest: none declared.

REFERENCES

Abiko,Y. *et al.* (2011) Participation of covalent modification of Keap1 in the activation of Nrf2 by tert-butylbenzoquinone, an electrophilic metabolite of butylated hydroxyanisole. *Toxicol. Appl. Pharmacol.*, **255**, 32–39.

Franceschini,A. *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.

Fullwood,M.J. *et al.* (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, **462**, 58–64.

Gil-Izquierdo,A. *et al.* (2012) Soy isoflavones and cardiovascular disease epidemiological, clinical and -omics perspectives. *Curr. Pharm. Biotechnol.*, **13**, 624–631.

Gu,J. and Liu,J.S. (2008) Bayesian biclustering of gene expression data. *BMC Genomics*, **9** (**Suppl 1**), S4.

Halliwell,B. *et al.* (2005) Health promotion by flavonoids, tocopherols, tocotrienols, and other phenols: direct or indirect effects? Antioxidant or not? *Am. J. Clin. Nutr.*, **81**, 268S–276S.

Hochreiter,S. *et al.* (2010) FABIA: factor analysis for bicluster acquisition. *Bioinformatics*, **26**, 1520–1527.

Huang,D.W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Huang,J. *et al.* (2013) Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network. *PLoS Comput. Biol.*, **9**, e1002998.

Iorio,F. *et al.* (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl Acad. Sci. USA*, **107**, 14621–14626.

Iskar,M. *et al.* (2013) Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. *Mol. Syst. Biol.*, **9**, 662.

Kaidery,N.A. *et al.* (2013) Targeting Nrf2-mediated gene transcription by extremely potent synthetic triterpenoids attenuate dopaminergic neurotoxicity in the MPTP mouse model of Parkinson's disease. *Antioxid Redox Signal.*, **18**, 139–157.

Kuiper,G.G. *et al.* (1998) Interaction of estrogenic chemicals and phytoestrogens with estrogen receptor beta. *Endocrinology*, **139**, 4252–4263.

Lamb,J. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Li,G. *et al.* (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, **148**, 84–98.

Messina,M. *et al.* (2009) Report on the 8th International Symposium on the Role of Soy in Health Promotion and Chronic Disease Prevention and Treatment. *J. Nutr.*, **139**, 796S–802S.

O'Leary,K.A. *et al.* (2004) Effect of flavonoids and vitamin E on cyclooxygenase-2 (COX-2) transcription. *Mutat Res.*, **551**, 245–254.

Okoh,V.O. *et al.* (2013) Reactive oxygen species via redox signaling to PI3K/AKT pathway contribute to the malignant growth of 4-hydroxy estradiol-transformed mammary epithelial cells. *PLoS One*, **8**, e54206.

Sandhu,K.S. *et al.* (2012) Large-scale functional organization of long-range chromatin interaction networks. *Cell Rep.*, **2**, 1207–1219.

Suthram,S. *et al.* (2010) Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Comput. Biol.*, **6**, e1000662.

Tham,D.M. *et al.* (1998) Clinical review 97: Potential health benefits of dietary phytoestrogens: a review of the clinical, epidemiological, and mechanistic evidence. *J. Clin. Endocrinol. Metab.*, **83**, 2223–2235.

Theocharis,S. *et al.* (2004) Peroxisome proliferator-activated receptor-gamma ligands as cell-cycle modulators. *Cancer Treat. Rev.*, **30**, 545–554.

Villa,A. *et al.* (2012) Tetradian oscillation of estrogen receptor alpha is necessary to prevent liver lipid deposition. *Proc. Natl Acad. Sci. USA*, **109**, 11806–11811.

Wishart,D.S. (2008) DrugBank and its relevance to pharmacogenomics. *Pharmacogenomics*, **9**, 1155–1162.

Yu,C.Y. *et al.* (2010) Genome-wide analysis of glucocorticoid receptor binding regions in adipocytes reveal gene network involved in triglyceride homeostasis. *PLoS One*, **5**, e15188.

Zhang,H.Y. *et al.* (2010) Evolutionary inspirations for drug discovery. *Trends Pharmacol. Sci.*, **31**, 443–448.

Zhang,T. *et al.* (2013) Activation of nuclear factor erythroid 2-related factor 2 and PPARgamma plays a role in the genistein-mediated attenuation of oxidative stress-induced endothelial cell injury. *Br. J. Nutr.*, **109**, 223–235.

Zhu,F. *et al.* (2012) Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res.*, **40**, D1128–D1136.