Genomics 94 (2009) 94–100

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Finding disease-specific coordinated functions by multi-function genes: Insight into the coordination mechanisms in diseases

Wencai Ma [a], Da Yang [a], Yunyan Gu [a], Xinwu Guo [a], Wenyuan Zhao [a], Zheng Guo [a,b,*]

[a] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150086, China
[b] Bioinformatics Centre, School of Life Science, University of Electronic Science and Technology of China, Chengdu, 610054, China

ARTICLE INFO

Article history:
Received 19 February 2009
Accepted 4 May 2009
Available online 8 May 2009

Keywords:
Disease gene
Gene functions
Gene ontology
Cancer
Gene mutation
Protein–protein interaction

ABSTRACT

We developed an approach using multi-function disease genes to find function pairs whose co-deregulation might induce a disease. Analyzing cancer genes, we found many cancer-specific coordinated function pairs co-deregulated by dysfunction of multi-function genes and other molecular changes in cancer. Studying two subtypes of cardiomyopathy, we found they show certain consistency at the functional coordination level. Our approach can also provide important information for finding novel disease genes as well as their mechanisms in diseases.

Introduction

Generally, the causation of a complex disease involves joint deregulation of multiple biological processes, instead of isolated disruption of some processes [1]. Therefore, understanding the mechanisms of functions coordinately contributing to a disease is important for studying the disease. In traditional functional genomic researches, usually based on Gene Ontology (GO) [2], a list of interesting genes for a disease are often translated into functional modules (categories enriched with the interesting genes), by tools such as DAVID [3], GO-2D [4] and others [5,6]. However, most of these tools interpret the interesting genes by finding isolated functional modules, providing no further information about their cooperative effects.

Recently, based on different assumptions, many computational methods have been designed for studying coordination (also termed interaction, crosstalk or interplay) relations of functions. Usually, two functions are defined as a coordinated function pair if they are densely connected by some relations such as co-expressions of genes [7], physical or genetic interactions of proteins [8,9] or others [10]. However, most current methods fail to exploit one interesting and important type of functional coordination regulated by multi-function genes which exist universally in biological systems. Obviously, the dysfunction of such multi-function genes may lead to co-deregulation of multiple functions responsible for some diseases [11]. For example, *CDKN1B* (cyclin-dependent kinase inhibitor 1B) possesses multiple functions such as differentiation, apoptosis and cell–cell adhesion [11] and its dysfunction may deregulate these functions, leading to cancer [11]. As another example, *APC* (adenomatosis polyposis coli) plays roles in several fundamental cellular processes, including cell adhesion and migration, organization of the actin and microtubule networks, spindle formation and chromosome segregation. Deregulation of these processes caused by mutations in *APC* is implicated in the initiation and expansion of colon cancer [12].

It is reasonable to assume that if the deregulation of two functions is responsible for inducing a disease, then the genes simultaneously involved in both functions may be more likely to be disease genes, which can be used as a clue for finding function pairs coordinately contributing to the disease. Here, in this paper, we proposed an approach to find such function pairs for a complex disease using multi-function disease genes. By analyzing a list of highly reliable cancer genes, we found many function pairs whose co-deregulation might be responsible for cancer. Then, based on protein–protein interaction [2] data and cancer genome mutation data, we found strong statistical evidences supporting that the coordinated function pairs identified by multi-function genes are highly likely to be co-deregulated by other molecular changes in cancer, such as dysfunction of pivot genes [9] or co-mutation of genes between functions. Besides suggesting new hints for the heterogeneous mechanisms of cancer, such statistically detectable evidences strongly support that the obtained function pairs may truly be co-deregulated in cancer. Finally, to show the proposed method can also be used to study other diseases,

* Corresponding author. College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150086, China. Fax: +86 28 83207187.
E-mail address: guoz@ems.hrbmu.edu.cn (Z. Guo).

0888-7543/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2009.05.001

we identified coordinated function pairs separately for two subtypes of cardiomyopathy and found a certain consistency of the two subtypes at the functional coordination level.

Results

CO-function pairs identified by multi-function disease genes for cancer

As demonstrated in Figs. 1A and B, we defined two functions as a coordinated function pair (referred to as a CO-function pair for short) if the disease gene ratio in their overlapped genes is significantly higher than the disease gene ratio in either of the two functions. By analyzing a list of cancer genes obtained from the Census database [13], with 10% FDR level (see Materials and methods), we found 86 CO-function pairs involving 61 functions for cancer (see Supplementary Table S1). For example, "cell cycle checkpoint" and "DNA repair" were identified as a CO-function pair because the ratio of cancer genes increased to 50% in their overlapped gene set, from 23% and 18% respectively in the two functions. Fig. 1C shows that, for each CO-function pair, the overlapped genes with multiple functions are much more likely to be cancer genes than the other genes in the two functions. We highlight that our method is able to discover some

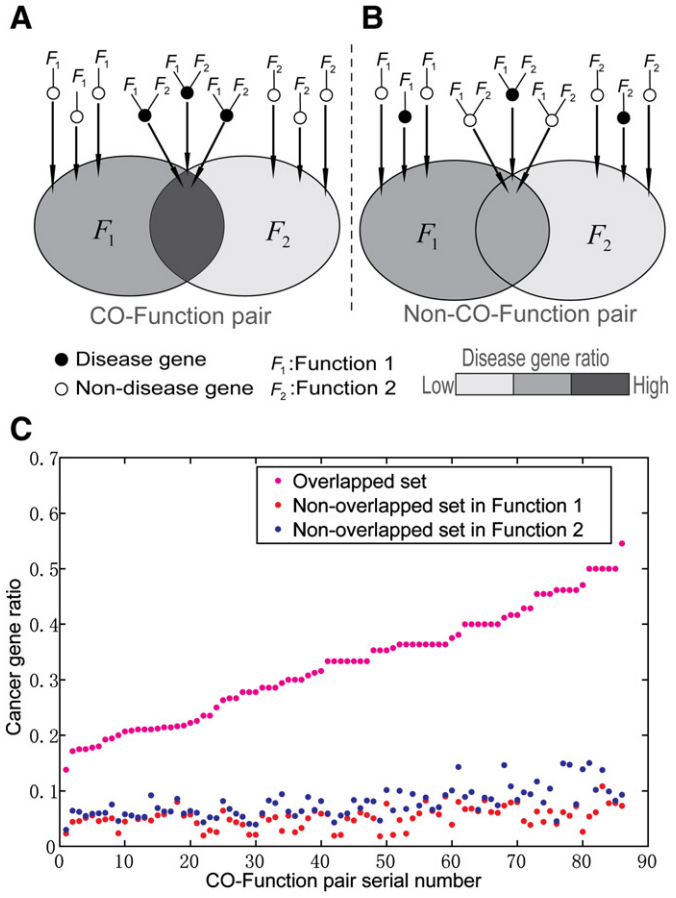

Fig. 1. CO-function pairs with increased disease gene ratios. (A) A CO-function pair. Genes with both functions in a CO-function pair (F_1 and F_2) may be more likely to be disease genes, and thus the disease gene ratio in the overlapped set will increase significantly. (B) A non CO-function pair. If F_1 and F_2 play independently for a disease, then the disease gene ratio in their overlapped genes will not increase. (C) Cancer gene ratio distributions in CO-function pairs. Along the transverse axis, the 86 CO-function pairs for cancer are ranked according to the ascending orders of the cancer gene ratios in theirs overlapped sets. For each point in the transverse axis, three dots with different colors represent the overlapped part and non-overlapped part of the two functions of a CO-function pair, respectively. It shows that for each pair, genes in the overlapped set contribute more to cancer.

important functions, though not enriched with cancer genes in general, contributing to cancer by mainly a subset of multi-function genes while the other genes in the functions are less relevant to the disease. For example, though "cell adhesion" is not enriched with cancer genes ($p = 0.66$ by hypergeometric test), it may contribute greatly to cancer by cooperating with other functions such as "negative regulation of cell proliferation". Notably, the cancer gene ratio is only 3% in "cell adhesion" and 8% in "negative regulation of cell proliferation". However, the cancer gene ratio in the genes with both functions is as high as 35%, indicating this subset of genes may play important roles in cancer through these two functions. On the contrary, the conventional enrichment methods neglect the fact that genes in the same function are heterogeneous in contributing to disease and thus cannot detect the special roles of multi-function genes.

Then, we took two multi-function genes, *ATM* (ataxia telangiectasia mutated) and *P53* (TP53, tumor protein p53), as examples to illustrate how multi-function genes play functional coordinator roles in cancer. *ATM* was found to be involved in three CO-function pairs: j "response to radiation" and "DNA repair"; k "response to abiotic stimulus" and "cell cycle"; and l "DNA repair" and "cell cycle checkpoint". As shown in Fig. 2A, in response to radiation or other abiotic stimulus that lead to DNA damage, ATM actives CHK2 (CHEK2) by phosphorylation [14,15], subsequently stimulating P53 signaling pathway to initiate cell cycle arrest and DNA repair which are important in preventing cancer when DNA damage happens [16]. Losing the ability of transferring the DNA damage signal to P53 by dysfunction of ATM can lead to the defects of cell cycle arrest and DNA repairs, which subsequently lead to cancer [17]. As another example, *P53* was found to be involved in five CO-function pairs: j "response to DNA damage stimulus" and "induction of apoptosis", k "regulation of transcription from RNA polymerase II promoter" and "cell cycle process", l "regulation of transcription from RNA polymerase II promoter" and "DNA repair", m "regulation of transcription from RNA polymerase II promoter" and "induction of apoptosis", and n"DNA repair" and "regulation of apoptosis". These findings match well with the mechanisms described in the "KEGG: P53 signaling pathway" (Fig. 2B): in response to DNA damage stimulus, P53 is activated to transcribe its targets that mediate cell cycle, DNA repair and apoptosis. It is reported that the mutation hot spots of P53 are involved in DNA binding [18] which can lead to a defect in transcription activity [19] and subsequently deregulation of cell cycle arrest, DNA repair and apoptosis in many tumors [20]. Thus, by studying these CO-function pairs, we can get a clear insight that both *ATM* and *P53* may play as coordinators of "cell cycle progression", "DNA repair", "apoptosis", and "in response to DNA damage".

Finally, we note that the proposed method can also be applied to study the coordination between biological pathways derived from other data sources such as KEGG [21]. For example, based on KEGG, we found that "P53 signaling pathway" may play coordinately with most of the cancer pathways described in KEGG. Many other pathway pairs coordinately playing in cancer were also revealed, such as "Wnt signaling pathway" and "Jak-STAT signaling pathway". To facilitate using these results, all the obtained cancer coordinated KEGG pathway pairs are provided in Supplementary Table S2.

CO-function network of cancer

The CO-function network was constructed by connecting the two functions with an edge in each identified CO-function pair. By presenting the attributes of the CO-function pairs in the network, we can get a clear view of all the CO-function pairs and find the most interesting ones (see Fig. 3). In this network, the top three hubs most frequently linked with others are "cell adhesion", "regulation of transcription from RNA polymerase II promoter" and "negative regulation of cell proliferation". Furthermore, viewing from the

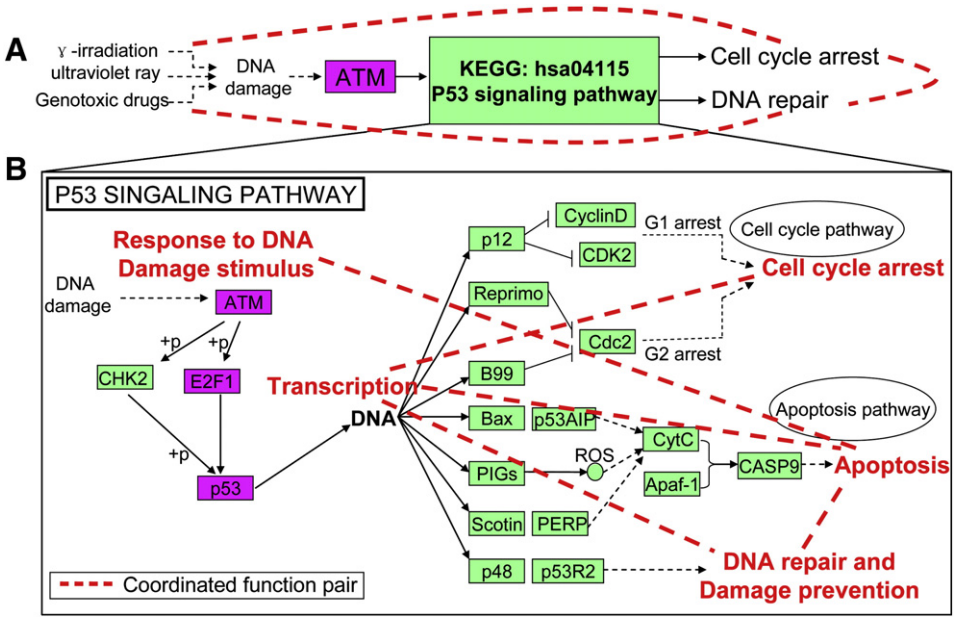

Fig. 2. CO-function pairs and multi-function genes explained by "KEGG:P53 signaling pathway". The red dashed lines mark some CO-function pairs found by our method. Genes colored by purple are discussed in detail in the text. (A) ATM plays its roles in inducing cancer by deregulating "cell cycle arrest" and "DNA repair" through the P53 signaling pathway. (B) P53 plays its multiple functions by regulating the transcription of different downstream genes.

thickness of the edges, we found that most of the CO-function pairs involving "cell adhesion" are highly significant, consolidating that "cell adhesion" coordinates with many other functions in contributing to cancer [22]. However, as mentioned above, "cell adhesion" is not enriched with cancer genes and will be undetectable by conventional enrichment analysis methods.

Then, we analyzed the top three most significant CO-function pairs (see Table 1 and Fig. 3).

(1) First, the most significant function pair is "cytoskeleton organization and biogenesis" and "negative regulation of cell proliferation" where the cancer gene ratios are 6% and 8% respectively, while the ratio in their overlapped genes increases to 46% (p-value 7.9×10^{-5}). The co-dysfunction of these two functions in inducing cancer was reported previously [23]. For example, mutations of *APC* can lead to the accumulation of β-catenin (CTNNB1) causing changes in "cytoskeletal organization" and "proliferation" responsible for cancer [23]. Thus, the deregulation of the coordination between "cytoskeleton organization" and "cell proliferation" in inducing cancer warrants further investigation.

(2) The second CO-function pair is "protein complex assembly" and "DNA metabolic process". The cancer gene ratios are 10% and 11% separately in these two functions, while the ratio in the overlapped set increases to 55% (p-value $=3.5\times10^{-4}$). "Protein complex assembly", including protein tetramerization, protein homooligomerization and docking [2], is critical for the activation of proteins. For example, tetramerization is the active conformation of P53 and is critically required for the stimulation of DNA damage [24]. Mutations in tetramerization domain causing the dysfunction of P53 can lead to inactive for DNA binding in DNA repair (a kind of "DNA metabolic process") and subsequently cancer [24].

(3) Finally, the third CO-function pair is "cell cycle process" and "regulation of transcription from RNA polymerase II promoter", in both of which the cancer gene ratio is 9%, while the ratio in their overlapped set increases to 46% (p-value $=4.0\times10^{-4}$). As mentioned above in the P53 signaling pathway, mutated *P53* gene may play its role in cancer by deregulating its transcriptional target genes affecting cell cycle process. Many other evidences suggesting the roles of these two coordinated functions in cancer can be found in [25].

Co-deregulation of CO-function pairs by different molecular changes in cancer

As discussed above, the CO-function pairs identified by multi-function genes may play important roles in cancer. Similarly to the effects of the dysfunction of multi-function genes, the co-deregulation of the obtained function pairs could also be carried by some other molecular changes, such as by the dysfunction of pivot genes densely linked to both functions [9] or the co-mutation of gene pairs between the two functions (see Fig. 4). For example, *E2F1* (E2F transcription factor 1) was found to be a pivot gene of the CO-function pair "response to radiation" and "regulation of apoptosis". As shown in Fig. 2, in response to the DNA damage induced by radiation, ATM activates E2F1 which can subsequently promote P53 induced apoptosis [15,26]. In addition, E2F1 may also play a role in the activation of ATM [26]. Thus, dysfunction of E2F1 may play a pivot role in co-deregulating the two functions, subsequently leading to cancer.

If the co-deregulation of two functions is truly responsible for cancer, then the pivot genes as coordinators of the two functions may be more likely to be cancer genes. Indeed, among the 118 distinct pivot genes identified for all the 86 CO-function pairs, 17 are known cancer genes which are significantly more than expected by random chance (hypergeometric test p-value $=5.9\times10^{-7}$). Next, we show that genes between the two functions of a CO-function pair tend to mutate together in cancer, which might be another way of co-deregulation of the obtained CO-function pairs in cancer. From all the 59,340 pairs formed by the 345 mutated genes in the cancer mutation data under this study, we found a total of 59 significantly co-mutated gene pairs in cancer samples (shown in Supplementary Table S3). In all the 59 co-mutated gene pairs, 14 pairs are between the functions in the CO-function pairs, which are significantly more than expected by random chance (hypergeometric test p-value $=0.02$). Thus, the found CO-function pairs tend to be co-deregulated by the co-mutation of gene pairs in cancer. The above results suggest that the co-deregulation of the CO-function pairs may truly contribute to cancer.

Fig. 3. The CO-function network of the CO-function pairs. The CO-function network was constructed by connecting every two functions in each of the identified CO-function pairs with an edge. Of all the 86 identified CO-function pairs, 83 pairs involving 56 functions (nodes) are connected together in the network. The node size is proportional to the degree of the node in the network and the color depth is proportional to the cancer gene ratio in this node. The thickness of an edge is proportional to the significance level (reciprocal of the *p*-value) of the coordination between the two functions linked by the edge. The nodes and edges indicated by red arrows are discussed in detail in the text.

Then, we analyzed an interesting CO-function pair "enzyme linked receptor protein signaling pathway" and "protein amino acid phosphorylation" that involves multi-function genes, pivot genes and co-mutated gene pairs. As shown in Fig. 4, the cancer gene ratios in the two functions are 9.6% and 7.0%, respectively, while the ratio in their overlapped set increases to 21% (p-value $= 1.9 \times 10^{-2}$).

Table 1
The three most significant CO-function pairs for cancer.

CO-function pair		Ratio	*p*-value
Cytoskeleton organization and biogenesis	Negative regulation of cell proliferation	(6%, 8%, 46%)[a] (6/13)[b]	7.9×10^{-5}
Protein complex assembly	DNA metabolic process	(10%, 11%, 55%)[a] (6/11)[b]	3.5×10^{-4}
Cell cycle process	Regulation of transcription from RNA polymerase II promoter	(9%, 9%, 46%)[a] (6/13)[b]	4.0×10^{-4}

[a] The three numbers are the cancer gene ratios in functions 1 and 2 and their overlapped gene set respectively.
[b] (C/A): "C" is the number of cancer genes in the overlapped gene set and "A" is the number of all the genes in the overlapped gene set.

There are 7 (21%) cancer genes in all the 33 multi-function genes, 6 (23%) cancer genes in the 26 pivot genes and 1 (50%) cancer gene in the 2 genes within 1 co-mutated gene pair. This high prevalence of cancer genes in multi-function genes, pivot genes and co-mutated genes of the CO-function may suggest to us new approaches of predicting cancer genes and providing their candidate cancer mechanisms for further experimental validation. To support this conjecture, for this CO-function pair, in addition to the above cancer genes in the Census database, we found 8 (24%) in the 33 multi-function genes, 5 (19%) in the 26 pivot genes and 1 (50%) in the 2 co-mutated genes that are defined as cancer genes in another four cancer gene databases [27–30]. Here, we termed these genes as candidate cancer genes (see Fig. 4). For example, *IGF1R* (insulin-like growth factor 1 receptor), possessing both functions of the CO-function pair, plays roles in mediating of the signaling pathways by regulating its downstream proteins [31], which subsequently regulate important cancer processes such as apoptosis, growth, proliferation and differentiation [31,32]. So, *IGF1R* may be a cancer gene whose dysfunction may induce cancer by reregulating this CO-function pair [33]. This example suggests the obtained CO-function pairs can provide valuable information for predicting cancer genes and simultaneously their mechanisms in inducing cancer.

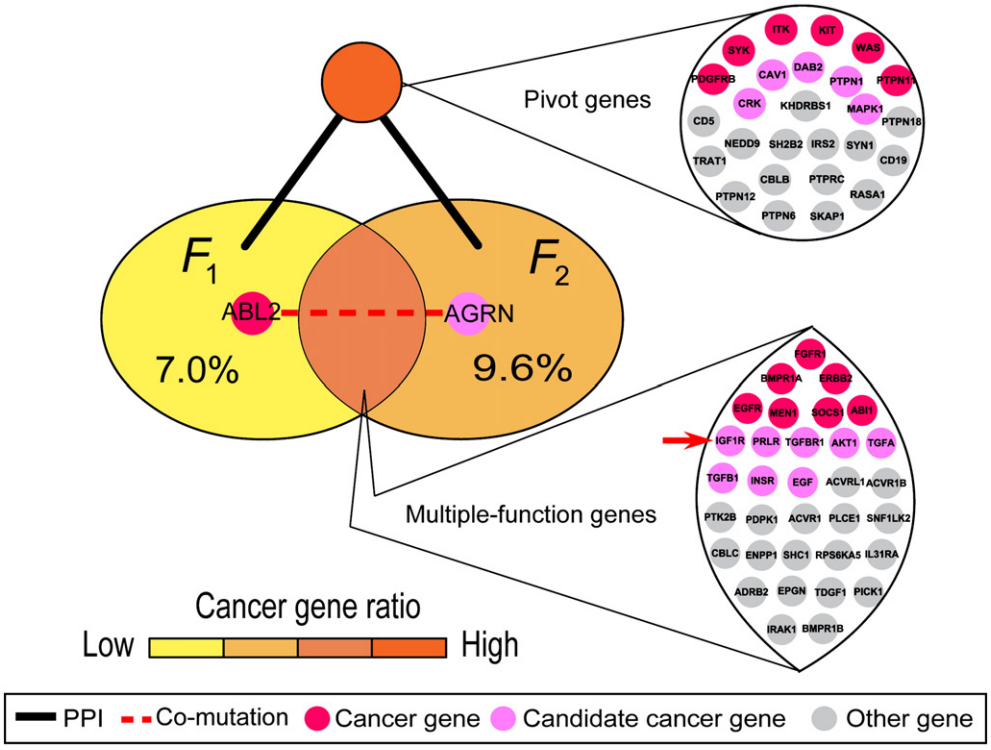

Fig. 4. Multi-function genes, pivot genes and co-mutated genes for a CO-function pair. F_1 and F_2 represent "protein amino acid phosphorylation" and "enzyme linked receptor protein signaling pathway", respectively. The color depth is proportional to the cancer gene ratio in F_1, F_2, the multi-function genes and the pivot genes. Note that cancer genes in Census and candidate cancer genes in other four cancer gene databases [27–30] are also discriminated by different colors. Pivot genes of F_1 and F_2 are indicated by PPI links to F_1 and F_2. Two co-mutated genes separately belong to F_1 and F_2 are linked by red dashed lines. *IGF1R* gene indicated by a red arrow is discussed in detail in the text.

CO-function pairs for cardiomyopathy

The proposed method can also be used to model coordinated functions for other complex diseases with multi-function disease genes. For illustration, we additionally studied and compared two subtypes of cardiomyopathy: dilated cardiomyopathy (DCM) and hypertrophic cardiomyopathy (HCM) with 18 and 15 disease genes (with GO "Biological Process" annotation) obtained from OMIM database [34], respectively. Using the same parameter setting as for cancer described above, we found 3 and 19 coordinated function pairs for DCM and HCM, respectively (see Supplementary Tables S4 and S5). Interestingly, although only 5 of the 18 DCM disease genes are shared by 15 HCM disease genes, all the 3 coordinated function pairs for DCM are shared by HCM. This result suggests a certain consistency of the two subtypes at the functional coordination level, supporting a previous report that the two distinct subtypes may share some key pathogenic mechanisms such as defects in the cardiac sarcomere [35].

As an example, we further analyzed a coordinated function pair "muscle development" and "muscle contraction" shared by DCM and HCM. Although DCM and HCM both involve dysfunction of this CO-function pair, the divergence of DCM and HCM is caused by affecting different aspects of the two functions. It has been suggested that impaired force transmission from the sarcomere to the surrounding syncytium (muscle contraction) may predispose affected myocytes to mechanical injury and cumulative cell death, secondary interstitial fibrosis, and cardiac dilation (muscle development), resulting in DCM [36–38]. On the other hand, HCM may be caused by the defects in force generation and impaired contractile performance (muscle contraction), which may lead to HCM through the compensatory hypertrophy remodeling of the heart (muscle development) [37,38]. We note that knowledge about the molecular mechanisms that trigger changes in cardiomyopathy is currently deficient [37], and thus many prominent functional coordination relations identified for it may warrant further biological investigations.

Discussion

The phenomenon that functions play coordinately in biological systems has been revealed by many wet-lab experiments [39]. In this paper, we proposed an approach to find coordinated function pairs involving multi-function disease genes. By analyzing a list of cancer genes based on GO or KEGG, we found many function pairs whose co-deregulation may induce cancer. Additionally, we found that the obtained CO-function pairs are highly likely to be co-deregulated in cancer by other molecular changes, including the dysfunction of pivot genes or co-mutation of genes between the two functions in each pair. This result suggests new hints for the heterogeneous mechanisms of cancer. As illustrated in the Results section, the multi-function genes, pivot genes and co-mutated genes, involved in the obtained CO-function pairs, can provide valuable information for finding novel cancer genes and their mechanisms in cancer. We additionally studied two subtypes of cardiomyopathy and identified some prominent disease-specific coordinated functions, showing that the proposed method can also be used to study other diseases.

For a given disease gene list, most current bioinformatics tools designed for finding functional modules enriched with the disease genes provide no information about functional coordination. Moreover, our proposed method can find some important disease-related pathways not enriched with disease genes, while such functions are undetectable by traditional enrichment analysis methods. For example, we found that "cell adhesion" is not enriched with cancer genes but it is the function most frequently cooperating with other functions in cancer. A resent study also reported that some causal functions of a disease are not necessarily enriched with the disease genes [40]. Notably, many functions in

the obtained CO-function pairs are common and typical cancer-related pathways. However, this does not mean the obtained CO-function pairs are too general to gain insights into the biological mechanisms of cancer. Oppositely, our findings highlight the collaboration between the pathways, rather than the individual pathways. To our knowledge, many prominent CO-function pairs obtained for the two diseases under this study have not been fully investigated and they would provide important hints for designing biological experiments to study the disease mechanisms at the functional collaboration level.

Our proposed method is specifically designed for finding coordinated function pairs involving multi-function disease genes. Obviously, it is important to develop this method to find other types of coordinated function pairs contributing to diseases. A possible approach for this extension is to integrate some functional links such as gene co-expression [7], protein physical and genetic interactions [8,9] and others [10], which warrant our future researches for revealing disease mechanisms.

Materials and methods

Data description

Based on the GO "Biological Process" ontology (released March 28, 2008) [2], we analyzed functions annotated with 15 to 500 genes, excluding the low-qualified annotation data denoted as IEA (Inferred from Electronic Annotation) and ND (No biological Data available) [41]. We also analyzed the pathways defined in the KEGG database [21] (downloaded in Dec. 1, 2008), which includes 214 signal and metabolic pathways.

The cancer genes were derived from the Cancer Gene Census database (referred to as the Census database for short) [13], which is the most frequently used cancer gene database with high quality. For a total of 367 cancer genes in the Census database, 261 with GO "Biological Process" annotation were analyzed in this study. The 15 and 18 disease genes for HCM and DCM respectively were obtained from the OMIM database [34].

The gene mutation data was integrated from four genome-wide somatic mutation screen datasets for a total of 2905 genes in 125 tumor samples of 4 cancer types [42–44]. We filtered out those genes mutated in only one sample and analyzed the remaining 709 genes. Our dataset choice tends to bias the co-occurring mutation gene pairs shared by various cancers, rather than gene pairs for a specific cancer type.

The human PPI data was derived from the HPRD database (Release 6) [45]. After removing self-interactions, 34,560 interactions among 9261 distinct human proteins were remaining for analysis.

Finding CO-function pairs

If the co-deregulation of two functions is responsible for inducing a disease, then the genes with both functions may be more likely to be disease genes. Based on this assumption, we defined two functions (F_1 and F_2) as a coordinated function pair (referred to as a CO-function pair) if the disease gene ratio (r_0) in the overlapped gene set ($O = F_1 \cap F_2$) of the two functions is significantly higher than both of the disease gene ratios (r_{F_1} and r_{F_2}) in the two functions (demonstrated in Figs. 1A, B). Statistically, the null hypothesis (H_0) and the alternative hypothesis (H_1) are as follows

$$H_0 : r_O = \max(r_{F_1}, r_{F_2}), H_1 : r_O > \max(r_{F_1}, r_{F_2}).$$

Suppose that the disease gene ratio in the overlapped genes of a function pair is $r_0 = k/n$ where k is the number of disease genes among n overlapped genes of the two functions, and in one of the two functions which has the higher disease gene ratio, there are M disease genes among a total of N genes. Then, the probability (p-value) of obtaining the disease gene ratio $r_0 = k/n$ in the overlapped genes of the two functions by random chance was estimated by the hypergeomertic distribution model [46] as follows

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}.$$

Considering the multiple testing problem, we applied a re-sampling based YB-FDR control procedure proposed by Yekutieli and Benjamini [47] because the test statistics are highly correlated and the number of tests for all the combinations of two functions in GO is extremely large. We did not use the BH-FDR approach proposed by Benjamini and Hochberg [48] because it assumes that all the test statistics are independent [47,49], which will yield very conservative results [6,49] if the test statistics are correlated and the number of tests is very large. The detailed procedure of the computationally intensive re-sampling procedure for YB-FDR control used in this study is described in Supplementary methods. Under a given FDR threshold, if two functions with ancestor–offspring relationship were both paired with a function, then the pair with the ancestor function was removed.

Finding pivot proteins and co-mutated gene pairs

A protein was defined as a pivot protein for a CO-function pair if its PPIs with both the non-overlapping parts of the two functions are more frequent than expected by random chance [9]. The statistical significance was evaluated by a hypergeometric test ($p \leq 0.01$).

Then, for finding co-mutated gene pairs in cancer based on the mutation data of 709 genes in 125 cancer samples, we calculated the mutual information (MI) value [50] of a gene pair (i,j) as

$$MI(i,j) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p_i(x) p_j(y)}.$$

$p_i(x)$ or $p_j(y)$ is the probability of $X = x$ or $Y = y$. Here, $x = 1$ or 0 if gene i is mutated or not in a sample, and for $y = 1$ or 0 if gene j is mutated or not in a sample. $p(x,y)$ is the joint probability of $X = x$ and $Y = y$.

After computing the MI values for all the gene pairs, we constructed the null distribution of random MI values by computing the MI values for gene pairs from 1000 sets of random mutation data produced by independently permuting each gene's mutation status in the samples while keeping its mutation times unchanged. If a pair of genes have a significantly high MI value, controlled by FDR = 10% (see details in Supplementary methods), they were defined as a co-mutated gene pair.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (grant nos. 30370388 and 30670539).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2009.05.001.

References

[1] T. Kim, J. Yoon, H. Cho, W.B. Lee, J. Kim, Y.H. Song, S.N. Kim, J.H. Yoon, J. Kim-Ha, Y.J. Kim, Downregulation of lipopolysaccharide response in *Drosophila* by negative crosstalk between the AP1 and NF-kappaB signaling modules, Nat. Immunol. 6 (2) (2005) 211–218.

[2] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat. Genet. 25 (1) (2000) 25–29.
[3] G. Dennis Jr., B.T. Sherman, D.A. Hosack, J. Yang, W. Gao, H.C. Lane, R.A. Lempicki, DAVID: Database for Annotation, Visualization, and Integrated Discovery, Genome Biol. 4 (5) (2003) P3.
[4] J. Zhu, J. Wang, Z. Guo, M. Zhang, D. Yang, Y. Li, D. Wang, G. Xiao, GO-2D: identifying 2-dimensional cellular-localized functional modules in Gene Ontology, BMC Genomics 8 (2007) 30.
[5] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. U. S. A. 102 (43) (2005) 15545–15550.
[6] W.T. Barry, A.B. Nobel, F.A. Wright, Significance analysis of functional categories in gene expression studies: a structured permutation approach, Bioinformatics 21 (9) (2005) 1943–1949.
[7] P. Langfelder, S. Horvath, Eigengene networks for studying the relationships between co-expression modules, BMC Syst. Biol. 1 (2007) 54.
[8] S. Bandyopadhyay, R. Kelley, N.J. Krogan, T. Ideker, Functional maps of protein complexes from quantitative genetic interaction data, PLoS Comput. Biol. 4 (4) (2008) e1000065.
[9] I. Ulitsky, R. Shamir, Pathway redundancy and protein essentiality revealed in the *Saccharomyces cerevisiae* interaction networks, Mol. Syst. Biol. 3 (2007) 104.
[10] H.Y. Chuang, E. Lee, Y.T. Liu, D. Lee, T. Ideker, Network-based classification of breast cancer metastasis, Mol. Syst. Biol. 3 (2007) 140.
[11] A. Sgambato, A. Cittadini, B. Faraglia, I.B. Weinstein, Multiple functions of p27 (Kip1) and its alterations in tumor cells: a review, J. Cell Physiol. 183 (1) (2000) 18–27.
[12] K. Aoki, M.M. Taketo, Adenomatous polyposis coli (*APC*): a multi-functional tumor suppressor gene, J. Cell Sci. 120 (Pt. 19) (2007) 3327–3335.
[13] P.A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, M.R. Stratton, A census of human cancer genes, Nat. Rev. Cancer 4 (3) (2004) 177–183.
[14] W.C. Lin, F.T. Lin, J.R. Nevins, Selective induction of E2F1 in response to DNA damage, mediated by ATM-dependent phosphorylation, Genes Dev. 15 (14) (2001) 1833–1844.
[15] J. Stanelle, B.M. Putzer, E2F1-induced apoptosis: turning killers into therapeutics, Trends Mol. Med. 12 (4) (2006) 177–185.
[16] M.B. Kastan, J. Bartek, Cell-cycle checkpoints and cancer, Nature 432 (7015) (2004) 316–323.
[17] S.E. Morgan, M.B. Kastan, p53 and ATM: cell cycle, cell death, and cancer, Adv. Cancer Res. 71 (1997) 1–25.
[18] Y. Cho, S. Gorina, P.D. Jeffrey, N.P. Pavletich, Crystal structure of a p53 tumor suppressor–DNA complex: understanding tumorigenic mutations, Science 265 (5170) (1994) 346–355.
[19] G.S. Jimenez, M. Nister, J.M. Stommel, M. Beeche, E.A. Barcarse, X.Q. Zhang, S. O Gorman, G.M. Wahl, A transactivation-deficient mouse model provides insights into Trp53 regulation and function, Nat. Genet. 26 (2000) 37–43.
[20] L. Raycroft, H.Y. Wu, G. Lozano, Transcriptional activation by wild-type but not transforming mutants of the p53 anti-oncogene, Science 249 (4972) (1990) 1049–1051.
[21] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, M. Kanehisa, KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Res. 27 (1) (1999) 29–34.
[22] U. Cavallaro, G. Christofori, Cell adhesion and signalling by cadherins and Ig-CAMs in cancer, Nat. Rev. Cancer 4 (2) (2004) 118–132.
[23] I. Nathke, Relationship between the role of the adenomatous polyposis coli protein in colon cancer and its contribution to cytoskeletal regulation, Biochem. Soc. Trans. 33 (Pt. 4) (2005) 694–697.
[24] P. Chene, The role of tetramerization in p53 function, Oncogene 20 (21) (2001) 2611–2617.
[25] K. Yoshida, Y. Miki, Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage, Cancer Sci. 95 (11) (2004) 866–871.
[26] J.T. Powers, S.K. Hong, C.N. Mayhew, P.M. Rogers, E.S. Knudsen, D.G. Johnson, E2F1 uses the ATM signaling pathway to induce p53 and Chk2 phosphorylation and apoptosis 1 1 American Cancer Society (ES Knudsen) and NIH (grants CA98601, ES11047, ES07784, CA16672, and T32ESO7247), Mol. Cancer Res. 2 (4) (2004) 203–214.
[27] B. Vogelstein, K.W. Kinzler, Cancer genes and the pathways they control, Nat. Med. 10 (8) (2004) 789–799.
[28] Hill M: http://embryology.med.unsw.edu.au/DNA/DNA10.htm. 1999.
[29] R.A. Baasiri, S.R. Glasser, D.L. Steffen, D.A. Wheeler, The breast cancer gene database: a collaborative information resource, Oncogene 18 (56) (1999) 7958–7965.
[30] Y. Yang, L.M. Fu, TSGDB: a database system for tumor suppressor genes, Bioinformatics 19 (17) (2003) 2311–2312.
[31] F. Peruzzi, M. Prisco, M. Dews, P. Salomoni, E. Grassilli, G. Romano, B. Calabretta, R. Baserga, Multiple signaling pathways of the insulin-like growth factor 1 receptor in protection from apoptosis, Mol. Cell Biol. 19 (10) (1999) 7203–7215.
[32] J. Riedemann, V.M. Macaulay, IGF1R signalling and its inhibition, Endocr. Relat. Cancer 13 (Suppl. 1) (2006) S33–S43.
[33] J.L. Resnik, D.B. Reichart, K. Huey, N.J. Webster, B.L. Seely, Elevated insulin-like growth factor I receptor autophosphorylation and kinase activity in human breast cancer, Cancer Res. 58 (6) (1998) 1159–1164.
[34] A. Hamosh, A.F. Scott, J.S. Amberger, C.A. Bocchini, V.A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, Nucleic Acids Res. 33 (Database issue) (2005) D514–D517.
[35] M. Patterson, Human genetics: affairs of the heart, Nature Rev. Genet. 2 (2001) 86.
[36] W.M. Franz, O.J. Muller, H.A. Katus, Cardiomyopathies: from genetics to the prospect of treatment, Lancet 358 (9293) (2001) 1627–1637.
[37] F. Ahmad, J.G. Seidman, C.E. Seidman, The genetic basis for cardiac remodeling, Annu. Rev. Genomics Hum. Genet. 6 (2005) 185–216.
[38] J. Mogensen, I.C. Klausen, A.K. Pedersen, H. Egeblad, P. Bross, T.A. Kruse, N. Gregersen, P.S. Hansen, U. Baandrup, A.D. Borglum, Alpha-cardiac actin is a novel disease gene in familial hypertrophic cardiomyopathy, J. Clin. Invest. 103 (10) (1999) R39–R43.
[39] B.O. Williams, L. Remington, D.M. Albert, S. Mukai, R.T. Bronson, T. Jacks, Cooperative tumorigenic effects of germline mutations in Rb and p53, Nat. Genet. 7 (4) (1994) 480–484.
[40] L. Wang, F. Sun, T. Chen, Prioritizing functional modules mediating genetic perturbations and their phenotypic effects: a global strategy, Genome Biol. 9 (12) (2008) R174.
[41] S.Y. Rhee, V. Wood, K. Dolinski, S. Draghici, Use and misuse of the gene ontology annotations, Nat. Rev. Genet. 9 (7) (2008) 509–515.
[42] S. Jones, X. Zhang, D.W. Parsons, J.C.H. Lin, R.J. Leary, P. Angenendt, P. Mankoo, H. Carter, H. Kamiyama, A. Jimeno, Core signaling pathways in human pancreatic cancers revealed by global genomic analyses, Science's STKE 321 (5897) (2008) 1801.
[43] D.W. Parsons, S. Jones, X. Zhang, J.C.H. Lin, R.J. Leary, P. Angenendt, P. Mankoo, H. Carter, I. Siu, An integrated genomic analysis of human glioblastoma multiforme, Science 321 (5897) (2008) 1807.
[44] L.D. Wood, D.W. Parsons, S. Jones, J. Lin, T. Sjoblom, R.J. Leary, D. Shen, S.M. Boca, T. Barber, J. Ptak, The genomic landscapes of human breast and colorectal cancers, Science 318 (5853) (2007) 1108.
[45] G.R. Mishra, M. Suresh, K. Kumaran, N. Kannabiran, S. Suresh, P. Bala, K. Shivakumar, N. Anuradha, R. Reddy, T.M. Raghavan, et al., Human protein reference database—2006 update, Nucleic Acids Res. 34 (Database issue) (2006) D411–D414.
[46] Y. Xing, Q. Xu, C. Lee, Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains, FEBS Lett. 555 (3) (2003) 572–578.
[47] D. Yekutieli, Y. Benjamini, Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics, J. Stat. Plan. Inference 82 (1–2) (1999) 171–196.
[48] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. 57 (1) (1995) 289–300.
[49] S. Pounds, C. Cheng, Improving false discovery rate estimation, Bioinformatics 20 (11) (2004) 1737–1745.
[50] J.P. Pluim, J.B. Maintz, M.A. Viergever, Mutual-information-based registration of medical images: a survey, IEEE Trans. Med. Imaging 22 (8) (2003) 986–1004.