The Author(s) *BMC Medical Genomics* 2017, **10**(Suppl 1):26
DOI 10.1186/s12920-017-0265-2

BMC Medical Genomics

RESEARCH Open Access

A disease similarity matrix based on the uniqueness of shared genes

Matthew B. Carson[1], Cong Liu[2,3], Yao Lu[3], Caiyan Jia[4] and Hui Lu[2,3,5*]

From The 6th Translational Bioinformatics Conference
Je Ju Island, Korea. 15-17 October 2016

Abstract

Background: Complex diseases involve many genes, and these genes are often associated with several different illnesses. Disease similarity measurement can be based on shared genotype or phenotype. Quantifying relationships between genes can reveal previously unknown connections and form a reference base for therapy development and drug repurposing.

Methods: Here we introduce a method to measure disease similarity that incorporates the uniqueness of shared genes. For each disease pair, we calculated the uniqueness score and constructed disease similarity matrices using OMIM and Disease Ontology annotation.

Results: Using the Disease Ontology-based matrix, we identified several interesting connections between cancer and other disease and conditions such as malaria, along with studies to support our findings. We also found several high scoring pairwise relationships for which there was little or no literature support, highlighting potentially interesting connections warranting additional study.

Conclusions: We developed a co-occurrence matrix based on gene uniqueness to examine the relationships between diseases from OMIM and DORIF data. Our similarity matrix can be used to identify potential disease relationships and to motivate further studies investigating the causal mechanisms in diseases.

Keywords: Disease-disease similarity, Disease-related genes, Clustering

Background

Over the last two decades computational methods have contributed increasingly to the analysis of many diseases [1, 2]. Areas of interest include the identification and annotation of disease genes [3–5], effects of single nucleotide polymorphisms (SNPs) [6], studies on gene-drug interactions [7], semantics and ontological work [8, 9], protein interaction networks [10], and many others. Of particular interest is the investigation of the relationship between diseases in terms of genotypic and phenotypic similarity. Recent work with disease networks has revealed the interconnected nature of various diseases [11, 12], which begs the question; can we gain new knowledge of a disease such as cancer by studying "connected", non-cancer diseases? Many diseases including obesity [13, 14], infection [15], diabetes [16], and possibly even psychological stress [17] have reported some relationship to cancer. Often the relationship type is unknown or partially known, indicating the need for further exploration of the interconnectedness of diseases. The key to understanding disease-disease similarity is to enrich the relationships with a quantifiable value and to infer new disease associations based on this enriched value.

Several strategies to measure disease similarity have been developed in previous studies. Mathur and Dinakarpandian used semantic similarity between ontological terms associated with diseases [18]. Using formal concept analysis (FCA, closely related to biclustering or co-clustering), Keller and colleagues

* Correspondence: huilu@uic.edu
[2]Department of Bioengineering, University of Illinois at Chicago, 851 S Morgan St, Chicago, IL 60607, USA
[3]Center for Biomedical Informatics, Shanghai Children's Hospital, 24 W Beijing Rd, Suite 1400, Shanghai 200000, China
Full list of author information is available at the end of the article

identified clusters from the previous known gene-disease associations [19]. By investigating formal concepts they revealed hidden relationships between diseases based on common associated genes as well as genes associated with a common set of diseases. Suthram et al. integrated high-throughput mRNA expression data and protein-protein interaction networks to discover human disease relationships in a systematic and quantitative way. They revealed similarities between diseases by identifying functional modules among the protein-protein interactions and scoring their association with diseases.

An alternative way to define disease similarity is to not only to consider the number of the genes they shared, but also to take into account the uniqueness of shared genes or molecular features. In this study, we introduce a method for measuring similarity between pairs of diseases based on the number of genes they share only with each other. We assume that if a gene or set of genes is related to only one pair of diseases, the similarity between those two diseases should be higher than that of a pair of disease sharing gene associations with many other diseases.

Methods

To analyze disease relationships, we built a disease co-occurrence matrix based on shared genes between each pair of diseases. We first calculated the *uniqueness* of each gene i as follows:

$$u_i = 1 - \sqrt{\frac{d_i}{d_n}}$$

where d_i is the number of diseases associated with each gene i and d_n is the number of diseases in the data set. Note that the fewer number of diseases related to a gene, the higher the possible uniqueness score for that gene.

Next, we created an $N \times N$ matrix. For each pair of diseases we added the uniqueness score of each shared gene:

$$d_{ij} = u_{s_1} + u_{s_2} + \ldots + u_{s_n}$$

where d_{ij} is a disease pair and u_{s_n} is the uniqueness value for each gene shared between the two. The diagonal elements of the disease co-occurrence matrix, where $i = j$ for d_{ij}, contain the sum of the uniqueness values for all genes related to disease d_i.

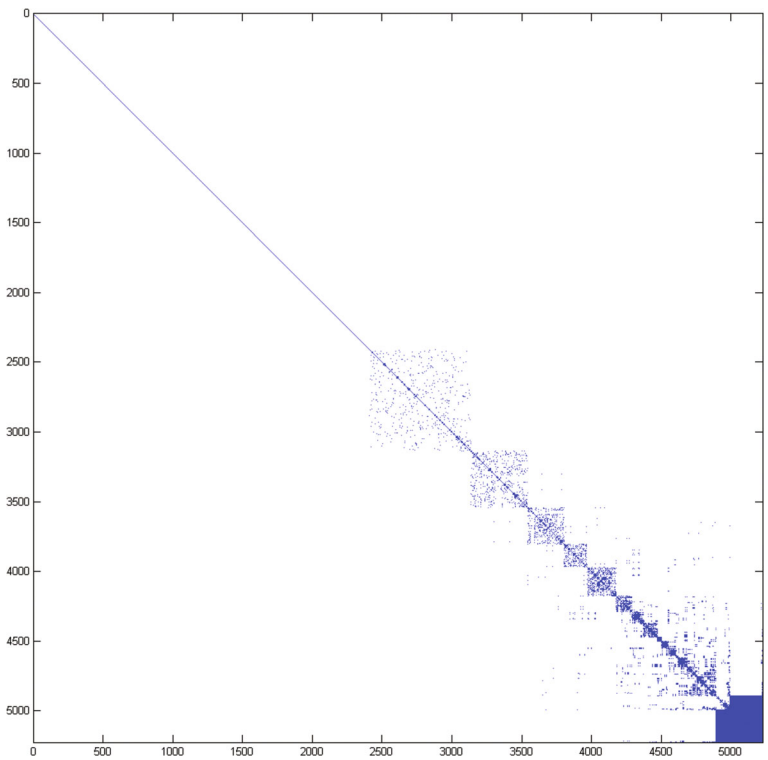

Fig. 1 A co-occurrence matrix showing the relationship between 5,224 diseases from the OMIM MorbidMap. Matrix elements colored blue indicate a relationship between two diseases, white elements indicate no relationship. Each blue matrix element (*i, j*) contains the sum of the uniqueness values for all genes related to both $disease_i$ and $disease_j$ (i.e. d_{ij}), while white elements are equal to 0. Diagonal elements indicate the identity relationship for each disease, i.e., the sum of the uniqueness values for all genes associated with $disease_i$. This figure was created using MATLAB [20]. The disease-gene relationships were extracted from OMIM MorbidMap

Next, we applied *symmetric approximate minimum degree permutation* to reorder the disease co-occurrence matrix. This algorithm was developed by Stefan I. Larimore and Timothy A. Davis and incorporated into MATLAB [20]. This reordering algorithm first creates a permutation vector p from a symmetric positive definite matrix A. This permutation vector, which contains a list of reordered columns from A, is then used to create a new matrix S such that $S = (p, p)$ has a sparser Cholesky factor than the original matrix A. The end result is that the reordered matrix S is less sparse near the lower diagonal and sparser near the upper diagonal. For our disease co-occurrence matrix, this effectively clusters highly related diseases in the lower right quadrant around the diagonal.

Results

We first applied this strategy to the OMIM MorbidMap database [21]. Fig. 1 shows the resulting reordered disease co-occurrence matrix for 5,224 diseases. While there are well-defined clusters, many of the cluster members are variations of the same disease phenotype or very closely related phenotypes. This is due to the high level of specificity of the OMIM disease categories. For example, the disease "46XY complete gonadal dysgenesis" is listed as two separate disease phenotypes, each with a different MIM identifier. While this distinction is important (the two phenotypes refer to mutations on different chromosomes), it does not serve our purposes in this case. We would like to see more relationships between phenotypically different diseases, and we would like very closely related phenotypes to be grouped together.

To address this, we created another matrix using gene-disease relationships gathered from Disease Ontology [22] and the GeneRIFs (Gene Reference Into Function) database (http://www.ncbi.nlm.nih.gov/gene/about-generif). The Disease Ontology provides a hierarchical structure in which more specific diseases can be grouped into broader categories, which allowed us to more easily compare phenotypically divergent diseases. These data sources were used by Osborne et al. to annotate the human genome [9] for disease (referred to hereafter as DORIF, http://projects.bioinformatics.northwestern.edu/do_rif/). The data set included 5,376 genes, 1,854 diseases, and 48,436 PubMed references relating genes to diseases. The DORIF co-occurrence matrix (Fig. 2) shows the comparisons between these diseases. There

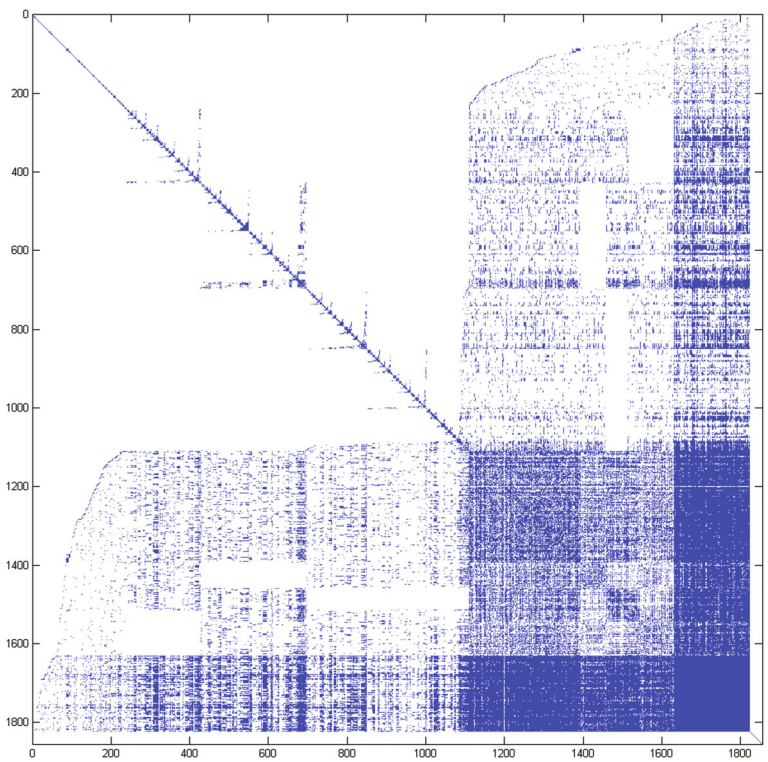

Fig. 2 A co-occurrence matrix showing the relationship between 1,854 diseases using DORIF data. Matrix elements colored blue indicate a relationship between two diseases, white elements indicate no relationship. Each blue matrix element (i, j) contains the sum of the uniqueness values for genes related to both $disease_i$ and $disease_j$ (i.e. d_{ij}), while white elements are equal to 0. Diagonal elements indicate the identity relationship for each disease, i.e., the sum of the uniqueness values for all genes associated with $disease_i$. This figure was created using MATLAB [20]. The disease-gene relationships were extracted from DORIF data

are two notable differences from the OMIM matrix. First, there are noticeably more disease relationships. This is because OMIM is a curated database of Medelian diseases, while DORIF is a 'Wiki-type' of resource (modifiable by NCBI users willing to provide their email address), leading to a higher depth and coverage of DORIF. In addition, given the denser disease-gene network in DORIF, we would expect to observe more disease relationships. Second, the DORIF matrix appears noisier; the relationships are not as tightly clustered as they are in the OMIM matrix. OMIM is a manually curated database and thus presumably has a higher accuracy rate. A slightly noisier matrix for DORIF in comparison to OMIM is not surprising. However, if the goal is to find hidden relationships between diseases, these "gray areas" are points of interest. A closer look at the individual clusters provides some interesting information.

Discussion

Figures 3 and 4 show a closer view of two different subsections of a dense cluster. Each of these figures is a 23 X 23 square submatrix of disease relationships from the solid blue cluster in the lower right-hand corner of the matrix in Fig. 2. The majority of the diseases in these submatrices are various types of cancers. There are some notable and interesting exceptions, however. For example, in the case of the relationship between malaria and cancer, the uniqueness value is close to that of those genes only related to malaria (0.47 and 0.57, respectively). Recent research provides some interesting findings about these two diseases. A clinical study showed that the mortality rate in patients with any type of cancer was increased after malarial infection [23]. Additionally, the malaria drug chloroquine has been shown to reduce tumor size in pancreatic cancer patients [24]. Another example is the relationship between hypertension induced by pregnancy and cancer. Recent work has shown that VEGF (Vascular endothelial growth factor) may be the connection. When taking anti-VEGF cancer drugs, patients develop very similar symptoms to pregnancy-induced hypertension. When VEGF expression levels are reduced in solid tumors, growth slows due to the lack of vascular development within. As a side effect, hypertensive symptoms occur [25]. Dental plaque and cancer appear to be highly related according to their uniqueness values as well; 0.77 (dental plaque) and 0.72 (dental plaque and cancer). Several past studies have made the connection between oral health and chronic illness. Recently, however, a clinical study spanning the last 24 years was released [26]. During this period, researchers followed 1,400 adults. They found that these subjects with high levels of dental plaque were 79% more likely to die prematurely from cancer. This work shows only an association between the two diseases, and thus the true nature of the relationship is yet

malaria	0.57	0.30	0.30	0.31	0.00	0.22	0.19	0.23	0.20	0.19	0.37	0.39	0.17	0.31	0.33	0.29	0.34	0.27	0.26	0.33	0.15	0.13	0.47
leukemia	0.30	1.09	0.52	0.50	0.35	0.34	0.28	0.19	0.22	0.38	0.77	0.46	0.27	0.31	0.29	0.37	0.48	0.31	0.45	0.32	0.40	0.25	0.95
acute leukemia	0.30	0.52	0.65	0.34	0.16	0.25	0.21	0.25	0.24	0.24	0.52	0.35	0.15	0.28	0.28	0.21	0.37	0.28	0.37	0.27	0.28	0.23	0.59
hypertension induced by pregnancy	0.31	0.50	0.34	0.88	0.35	0.34	0.26	0.41	0.29	0.41	0.61	0.49	0.40	0.35	0.38	0.35	0.52	0.35	0.43	0.35	0.14	0.34	0.72
Fibroid tumor	0.00	0.35	0.16	0.35	0.59	0.30	0.18	0.24	0.28	0.51	0.44	0.23	0.24	0.20	0.20	0.18	0.26	0.14	0.30	0.00	0.24	0.20	0.55
liver metastasis	0.22	0.34	0.25	0.34	0.30	0.63	0.14	0.25	0.16	0.38	0.52	0.31	0.25	0.22	0.33	0.24	0.35	0.19	0.16	0.27	0.32	0.24	0.62
autistic disorder	0.19	0.28	0.21	0.26	0.18	0.14	0.48	0.13	0.25	0.23	0.30	0.19	0.21	0.17	0.09	0.10	0.28	0.23	0.25	0.08	0.00	0.24	0.33
cardiovascular system disease	0.23	0.19	0.25	0.41	0.24	0.25	0.13	0.45	0.19	0.30	0.35	0.17	0.22	0.19	0.29	0.13	0.32	0.28	0.25	0.00	0.12	0.24	0.43
dementia	0.20	0.22	0.24	0.29	0.28	0.16	0.25	0.19	0.54	0.25	0.36	0.26	0.21	0.00	0.00	0.22	0.40	0.29	0.37	0.22	0.25	0.29	0.48
uterine Fibroids	0.19	0.38	0.24	0.41	0.51	0.38	0.23	0.30	0.25	0.64	0.52	0.27	0.28	0.29	0.28	0.18	0.34	0.27	0.30	0.00	0.17	0.20	0.61
malignant neoplasm of lung	0.37	0.77	0.52	0.61	0.44	0.52	0.30	0.35	0.36	0.52	1.21	0.54	0.40	0.47	0.41	0.30	0.56	0.33	0.54	0.42	0.48	0.33	1.10
Behcet syndrome	0.39	0.46	0.35	0.49	0.23	0.31	0.19	0.17	0.26	0.27	0.54	0.71	0.33	0.29	0.32	0.36	0.40	0.20	0.34	0.41	0.25	0.22	0.61
Fetal Growth Retardation	0.17	0.27	0.15	0.40	0.24	0.25	0.21	0.22	0.21	0.28	0.40	0.33	0.55	0.23	0.25	0.22	0.35	0.25	0.25	0.19	0.18	0.26	0.51
relapsing pancreatitis	0.31	0.31	0.28	0.35	0.20	0.22	0.17	0.19	0.00	0.29	0.47	0.29	0.23	0.53	0.25	0.17	0.34	0.27	0.30	0.23	0.14	0.19	0.50
temporal arteritis	0.33	0.29	0.28	0.38	0.20	0.33	0.09	0.29	0.00	0.28	0.41	0.32	0.25	0.25	0.46	0.28	0.37	0.19	0.20	0.29	0.10	0.00	0.45
mucocutaneous lymph node syndrome	0.29	0.37	0.21	0.35	0.18	0.24	0.10	0.13	0.22	0.18	0.30	0.36	0.22	0.17	0.28	0.50	0.31	0.15	0.22	0.25	0.15	0.10	0.41
Dental Plaque	0.34	0.48	0.37	0.52	0.26	0.35	0.28	0.32	0.40	0.34	0.56	0.40	0.35	0.34	0.37	0.31	0.77	0.34	0.47	0.39	0.29	0.34	0.72
Downs syndrome	0.27	0.31	0.28	0.35	0.14	0.19	0.23	0.28	0.29	0.27	0.33	0.20	0.25	0.27	0.19	0.15	0.34	0.56	0.30	0.07	0.17	0.31	0.49
Parkinson disease	0.26	0.45	0.37	0.43	0.30	0.16	0.25	0.25	0.37	0.30	0.54	0.34	0.25	0.30	0.20	0.22	0.47	0.30	0.72	0.28	0.24	0.35	0.59
cystic fibrosis	0.33	0.32	0.27	0.35	0.00	0.27	0.08	0.00	0.22	0.00	0.42	0.41	0.19	0.23	0.29	0.25	0.39	0.07	0.28	0.58	0.19	0.17	0.50
colon carcinoma	0.15	0.40	0.28	0.14	0.24	0.32	0.00	0.12	0.25	0.17	0.48	0.25	0.18	0.14	0.10	0.15	0.29	0.17	0.24	0.19	0.65	0.22	0.59
mental depression	0.13	0.25	0.23	0.34	0.20	0.24	0.24	0.24	0.29	0.20	0.33	0.22	0.26	0.19	0.00	0.10	0.34	0.31	0.35	0.17	0.22	0.57	0.47
cancer	0.47	0.95	0.59	0.72	0.55	0.62	0.33	0.43	0.48	0.61	1.10	0.61	0.51	0.50	0.45	0.41	0.72	0.49	0.59	0.50	0.59	0.47	1.77

Fig. 3 A subset of the disease co-occurrence matrix and the relationships between 23 diseases beginning with malaria (top) and ending with cancer (bottom). Disease labels for the rows apply to the columns as well. The value of each element (*i, j*)is the sum of the uniqueness values of all genes related to both $disease_i$ and $disease_j$ (i.e. d_{ij}). Darker squares indicate a higher uniqueness value. This figure was created using MATLAB [20]. The disease-gene relationships were extracted from DORIF data

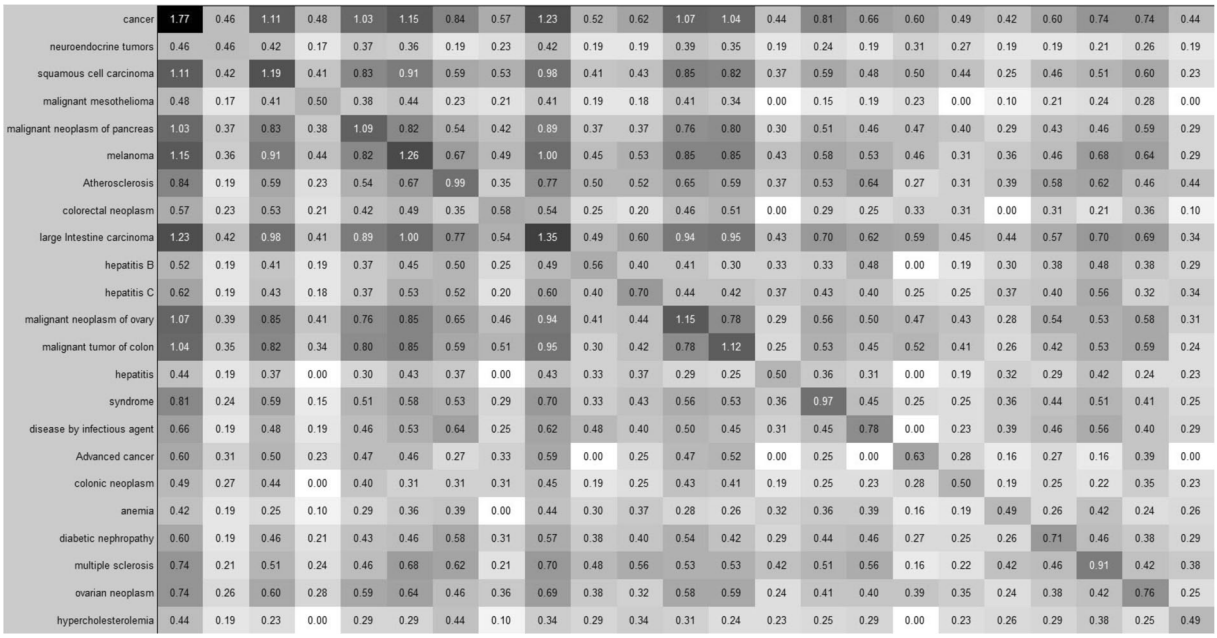

Fig. 4 A subset of the disease co-occurrence matrix and the relationships between 23 diseases beginning with cancer (top) and ending with hypercholesterolemia (bottom). Disease labels for the rows apply to the columns as well. The value of each element (*i, j*) is the sum of the uniqueness values of all genes related to both $disease_i$ and $disease_j$ (i.e. d_{ij}). Darker squares indicate a higher uniqueness value. This figure was created using MATLAB [20]. The disease-gene relationships were extracted from DORIF data

to be discovered. Other relationships from our matrix share a high uniqueness score, but there is little or no experimental evidence linking them. For example, migraine headaches and large intestine carcinoma have a shared uniqueness score of 0.403, while migraine alone is 0.498 (not shown in figures). Despite this, we could not find research references linking the two. However, this matrix could be used to identify potential disease relationships and to motivate further study into the elucidation of causal mechanisms in disease.

Conclusions

We developed a co-occurrence matrix based on gene uniqueness to examine the relationships between diseases from the OMIM and DORIF databases. We found examples of known disease relationships as well as connections with no available evidence. This matrix serves as a preliminary reference for identifying disease-disease associations, providing a map of the connections between diseases, and directing focus toward those associations which may not otherwise be obvious. It could also be used as a first step in drug repositioning research, directing focus to new potential protein or DNA targets. It is important to note that the purpose of this study is to provide a disease similarity matrix from the uniqueness of shared genes as a reference and that it is not meant to serve as the basis for clinical decisions in patient care.

Complex diseases such as cancer are both unique and related to other diseases, and analyzing all pairwise relationships between diseases provides new perspectives. For instance, drugs used for the treatment of related non-cancer diseases may help to treat the side effects of cancer drugs. Another example lies in the complex relationship between bacteria and cancer: bacteria can be both beneficial and cancer causing. Research on disease relationships can stimulate the development of new ideas about cancer and its relationship to infection. Additionally, this research could help clarify the mechanisms and tissue-specificity of non-cancer diseases and how they may prime the cellular environment for metastasis. We expect that in the near future, due to the availability of an enormous amount of genotypic and phenotypic data related to disease, there will be a novel view point for cancer research emerging from these studies.

Acknowledgements

Not applicable

Funding

Publication of this article was funded by UIC and by the National Natural Science Foundation of China (No.31071167 and No.31370751).

Availability of data and materials

OMIM data is available from https://www.omim.org/. DORIF data is available from http://projects.bioinformatics.northwestern.edu/do_rif/.

Authors' contributions
MC and CL: designed the concept, developed statistical methods, collected the real data, performed the clustering analysis, and drafted the manuscript. YL and CJ developed statistical methods, implemented the coding, and approved the final manuscript. HL designed the concept, provided financial support, and approved the final manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

About this supplement
This article has been published as part of BMC Medical Genomics Volume 10 Supplement 1, 2017: Selected articles from the 6th Translational Bioinformatics Conference (TBC 2016): medical genomics. The full contents of the supplement are available online at https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-10-supplement-1.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, 680 N Lake Shore Dr, Suite 1400, Chicago, IL 60611, USA. [2]Department of Bioengineering, University of Illinois at Chicago, 851 S Morgan St, Chicago, IL 60607, USA. [3]Center for Biomedical Informatics, Shanghai Children's Hospital, 24 W Beijing Rd, Suite 1400, Shanghai 200000, China. [4]Department of Computer Science, Beijing Jiaotong University, No.3 Shangyuancun, Haidian District, Beijing 100044, China. [5]SJTU-Yale Joint Center for Biostatistics, Department of Bioinformatics and Biostatistics, Shanghai Jiaotong University, 800 Dongchuan Road, Shanghai 200000, China.

Published: 24 May 2017

References
1. Gu JL, Lu Y, Liu C, Lu H. Multiclass classification of sarcomas using pathway based feature selection method. J Theor Biol. 2014;362:3–8.
2. Liu C, Xu J, Chen Y, Guo X, Zheng Y, et al. Characterization of genome-wide H3K27ac profiles reveals a distinct PM2.5-associated histone modification signature. Environ Health. 2015;14:65. doi:10.1186/s12940-015-0052-5..
3. Wang X, Gotoh O. Inference of cancer-specific gene regulatory networks using soft computing rules. Gene Regul Syst Bio. 2010;4:19–34.
4. Zheng B, Liu J, Gu J, Lu Y, Zhang W, et al. A three-gene panel that distinguishes benign from malignant thyroid nodules. Int J Cancer. 2015;136:1646–54.
5. Qin W, Liu C, Sodhi M, Lu H. Meta-analysis of sex differences in gene expression in schizophrenia. BMC Syst Biol. 2016;10 Suppl 1:9.
6. Li H, Lee Y, Chen JL, Rebman E, Li J, et al. Complex-disease networks of trait-associated single-nucleotide polymorphisms (SNPs) unveiled by information theory. J Am Med Inform Assoc. 2012;19:295–305.
7. Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, et al. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. PLoS Comput Bio. 2010;6:e1000662.
8. Osborne JD, Lin S, Kibbe WA, Zhu L, Danila MI, Chisholm RL. GeneRIF is a more comprehensive, current and computationally tractable source of gene-disease relationships than OMIM. Bioinformatics Core, Northwestern University Technical Report; 2007.
9. Osborne JD, Flatow J, Holko M, Lin SM, Kibbe WA, et al. Annotating the human genome with Disease Ontology. BMC Genomics. 2009;10 Suppl 1:S6.
10. Ideker T, Sharan R. Protein networks in disease. Genome Res. 2008;18:644–52.
11. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, et al. The human disease network. Proc Natl Acad Sci U S A. 2007;104:8685–90.
12. Zhang M, Zhu C, Jacomy A, Lu LJ, Jegga AG. The orphan disease networks. Am J Hum Genet. 2011;88:755–66.
13. Kushi LH, Byers T, Doyle C, Bandera EV, McCullough M, et al. American Cancer Society Guidelines on Nutrition and Physical Activity for cancer prevention: reducing the risk of cancer with healthy food choices and physical activity. CA Cancer J Clin. 2006;56:254–81. quiz 313-254.
14. Taubes G. Cancer research. Unraveling the obesity-cancer connection Science. 2012;335(28):30–22.
15. Anand P, Kunnumakkara AB, Sundaram C, Harikumar KB, Tharakan ST, et al. Cancer is a preventable disease that requires major lifestyle changes. Pharm Res. 2008;25:2097–116.
16. Wang F, Herrington M, Larsson J, Permert J. The relationship between diabetes and pancreatic cancer. Mol Cancer. 2003;2:4.
17. Garssen B. Psychological factors and cancer development: evidence after 30 years of research. Clin Psychol Rev. 2004;24:315–38.
18. Mathur S, Dinakarpandian D. Finding disease similarity based on implicit semantic similarity. J Biomed Inform. 2012;45:363–71.
19. Keller BJ, Eichinger F, Kretzler M. Formal concept analysis of disease similarity. AMIA Summits Transl Sci Proc. 2012;2012:42–51.
20. MATLAB. version 7.14.0 (R2012a). Natick: The MathWorks Inc; 2012.
21. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33:D514–517.
22. Schriml LM, Arze C, Nadendla S, Chang YW, Mazaitis M, et al. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. 2012;40:D940–946.
23. Lehrer S. Association between malaria incidence and all cancer mortality in fifty U.S. States and the District of Columbia. Anticancer Res. 2010;30:1371–3.
24. Yang S, Wang X, Contino G, Liesa M, Sahin E, et al. Pancreatic cancers require autophagy for tumor growth. Genes Dev. 2011;25:717–29.
25. Gluhovschi G, Gluhovschi A, Petrica L, Anastasiu D, Gluhovschi C, et al. Pregnancy-induced hypertension–a particular pathogenic model. Similarities with other forms of arterial hypertension. Rom J Intern Med. 2012;50:71–81.
26. Soder B, Yakob M, Meurman JH, Andersson LC, Soder PO. The association of dental plaque with cancer mortality in Sweden. A longitudinal study BMJ Open. 2012;2.