Consideration of Statistical vs. Biological Significances for Omics Data-Based Pathway Network Analysis

Xianquan Zhan [1, 2, 3, 4*], Ying Long [1, 2, 3], Xiaohan Zhan [1, 2, 3], Yun Mu [1, 2, 3]

[1] Key Laboratory of Cancer Proteomics of Chinese Ministry of Health, Xiangya Hospital, Central South University, 87 Xiangya Road, Changsha, Hunan 410008, P. R. China

[2] Hunan Engineering Laboratory for Structural Biology and Drug Design, Xiangya Hospital, Central South University, 87 Xiangya Road, Changsha, Hunan 410008, P. R. China

[3] State Local Joint Engineering Laboratory for Anticancer Drugs, Xiangya Hospital, Central South University, 87 Xiangya Road, Changsha, Hunan 410008, P. R. China

[4] The State Key Laboratory of Medical Genetics, Central South University, 88 Xiangya Road, Changsha, Hunan 410008, P. R. China

***Corresponding author**: Xianquan Zhan, Key Laboratory of Cancer Proteomics of Chinese Ministry of Health, Xiangya Hospital, Central South University, 87 Xiangya Road, Changsha, Hunan 410008, P.R. China. Tel: +86-731-48327905; Fax: +86-731-48327905; E-mail: yjzhan2011@gmail.com

http://mo.qingres.com

OPEN ACCESS

DOI: 10.20900/mo.20170002

Received: November 20, 2016

Accepted: January 10, 2017

Published: February 25, 2017

ABSTRACT

In the era of biological omics, an essential issue is how to mine the important biological information, including important signaling molecule patterns, signaling pathways, molecular networks, and pathway-network systems, from big omic data in combination with phenotype features in a given biological system. An appropriate statistical method and algorithm play central technique-support roles in such a bioinformation-mining process. However, for the statistical results, one must realize the difference and relationship of statistical vs. biological significances in those analysis processes. Statistical significance and biological significance are two different concepts with overlapping of their results. The choice of statistical method and threshold value of statistical significance should be determined by the data type and research goal. A statistically significant result must be reasonably interpreted with corresponding biological processes to decide its biological significance. One must not use statistical significance to kidnap biological significance, and the statistical result is only a reference to determine a biological significance.

Keywords: Omics data, Pathway network analysis, Statistical significance, Biological significance

1 INTRODUCTION

The rapid development of biological omics (genomics, transcriptomics, proteomics, peptidomics, and metabolomics) [1, 2] and systems biology [3-5] is driving personalized precision medicine [6]. Biological statistics plays a key role in this process to obtain important information from those complicated omics data and phenotype data [7-9]. However, one must realize the difference and relationship between statistical significance and biological significance. This article addresses the importance of biological omics and systems biology, necessity of an appropriate statistical method and algorithm for biological omics data analysis, a real example, and molecular network concept-based statistical consideration and biological significance.

2 IMPORTANCE OF BIOLOGICAL OMICS AND SYSTEMS BIOLOGY

The development of modern molecular medicine is experiencing at least three paradigm shifts (Fig. 1): (i) from macrocosmic view to microcosmic view, which is from anatomy, histology, cytobiolgy, molecular biology, to structural biology. (ii) From traditional single parameter strategy to multi-parameter systematic strategy [1, 2, 10, 11], because a traditional single-molecule biomarker is based on an unrealistic assumption that an increase in the amount of a single compound can unambiguously specify a disease [10]. Whereas, the reality is that cancer is a whole-body disease that alters in different levels of multiple genes, multiple proteins, and multiple metabolites, involves multiple causes, multiple processes, and multiple consequences; No a single molecule, or single signaling pathway is able to resolve all problems of a cancer. (iii) From the same treatment for the same types of disease to personalized/precise treatment, because personalized/individualized variations are involved in the each aspect of healthcare (Fig. 2) [1, 2]: prediction/prevention/assessment of preventive response, early-stage diagnosis/therapy/assessment of therapeutic response, and late-stage diagnosis/therapy/assessment of therapeutic response.

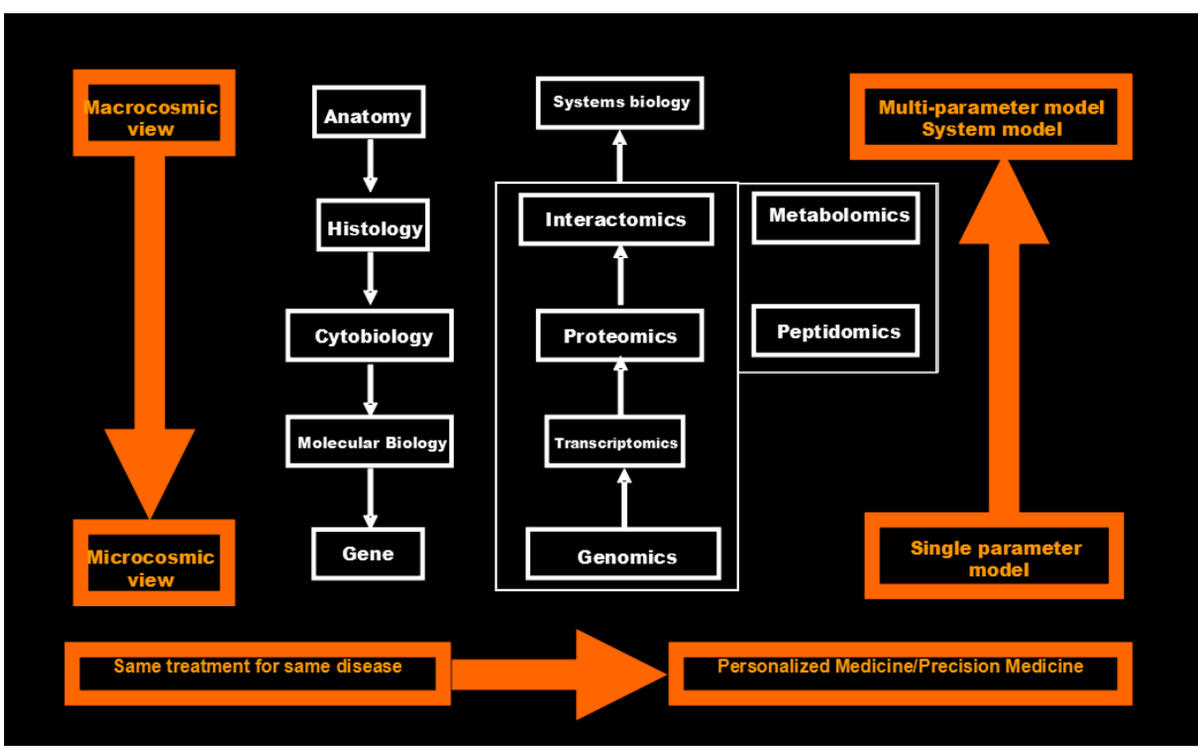

Fig. 1 Three paradigms shift in the field of modern oncology.

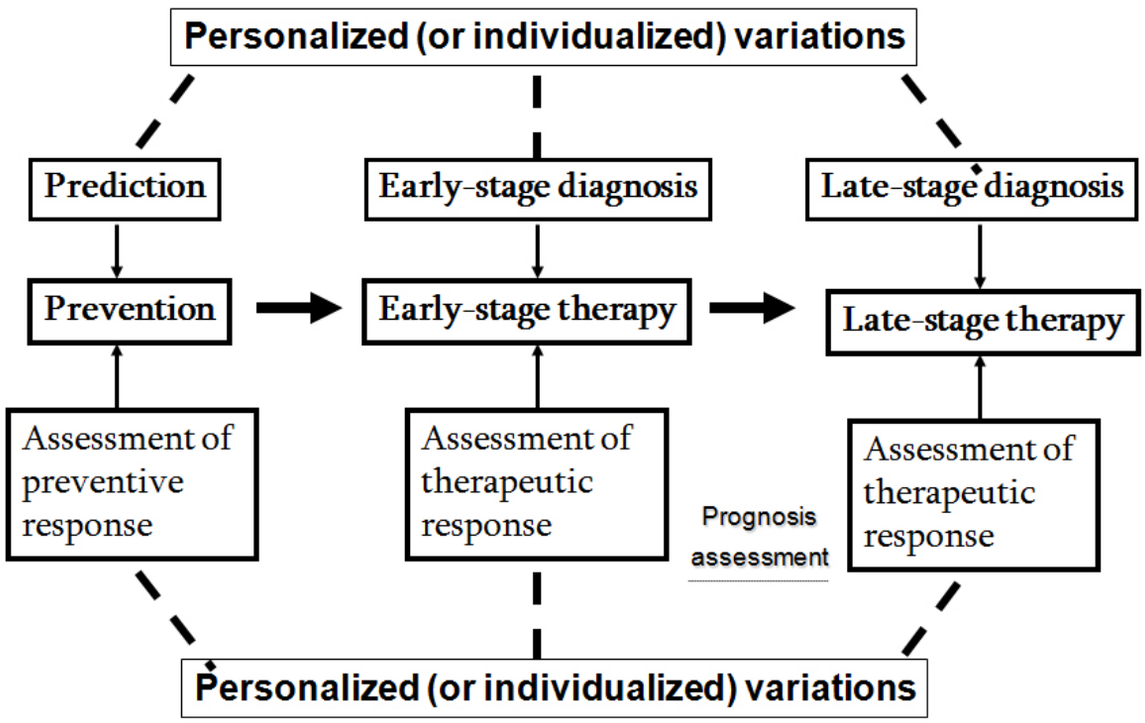

Fig. 2 Personalized variations are involved in each aspect of healthcare. Reproduced from Hu, Wang, and Zhan (2013) [2], with permission from BioMed Central Publisher. Copyright 2013 remains with authors due to the open-access article under the Creative Commons Attribution License.

The paradigm shifts (ii) and (iii) mainly benefit from the rapid developments of omics (genomics, transcriptomics, proteomics, peptidomics, metabolomics, and interactomics) and systems biology together with identification of phenotype including different clinical characteristics (Fig. 3) [2-5, 12, 13]. Furthermore, the genetic central rule (Fig. 4) [14-16] reveals that genome contains 23-pair chromosomes and about 25000 genes with two main approaches (gene sequencing and gene chip) to identify gene mutation, insert, loss, and fusion. Transcriptome contains coding RNAs (mRNAs) and non-coding RNAs (ncRNAs) with two main approaches (sequencing and microarray) [17] to identify variations of gene transcriptions, here splicing variations cause that one gene corresponds to multiple transcripts, which results in that transcriptome is much complicated than genome. Protein is synthesized in the ribosome with the guidance of each transcript. Each protein molecule has multiple copy numbers, and has to translocate to the corresponding location, form specific spatial conformation, and interact with the surrounding molecules, to exert its biological functions. Moreover, lots of post-translational modifications (PTMs) [18, 19] would occur in the processes from synthesized protein to specific location. Those factors, such as protein copy number, splicing, PTMs, translocation, and spatial conformation, result in that proteome is much more complicated than transcriptome and genome [19]. Metabolome contains all metabolites that are derived from sugars, lipids, proteins, and nucleic acids [20]. The variations in metabolome can reflect the mechanism of a biological process and the phenotype characteristics [21].

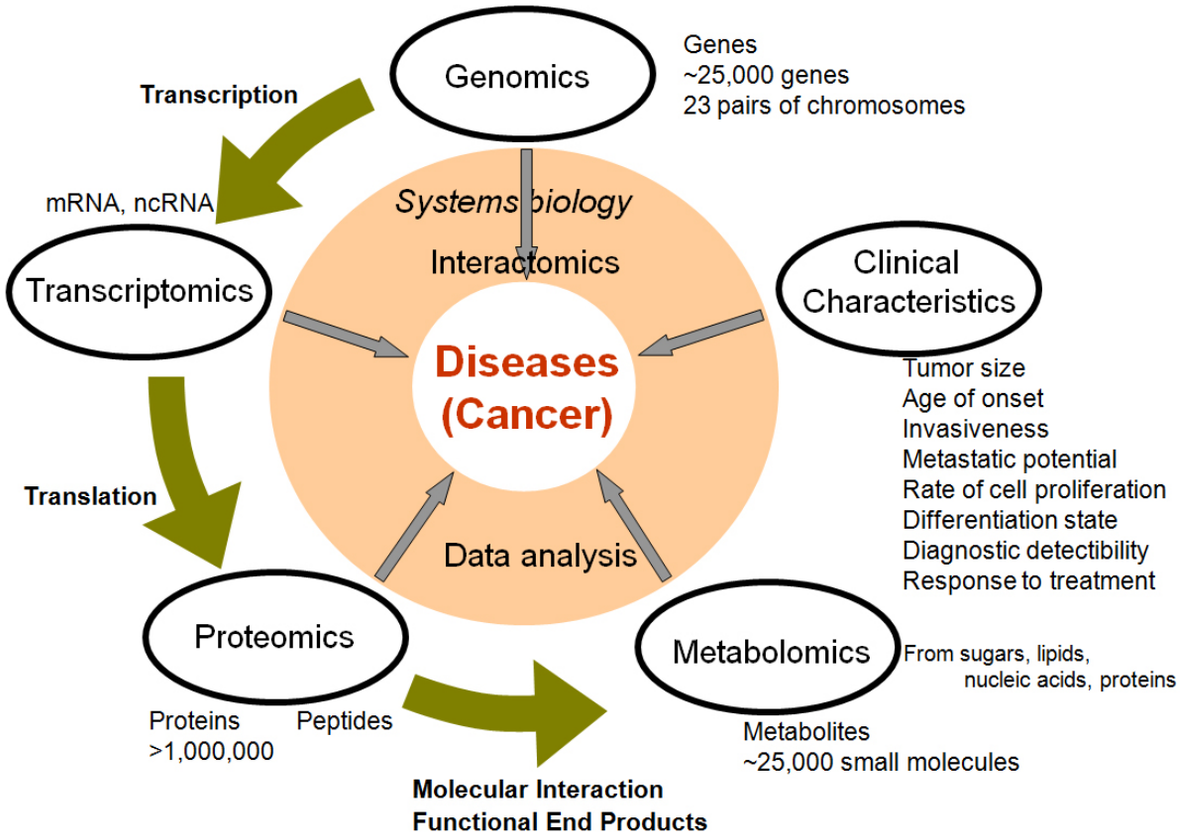

Fig. 3 Contributions of omics and systems biology to cancer biology. Modified from Hu, Wang, and Zhan (2013) [2] , with permission from BioMed Central Publisher. Copyright 2013 remains with authors due to the open-access article under the Creative Commons Attribution License.

3 NECESSITY OF AN APPRORIATE STATISTICAL METHOD FOR BIOLOGICAL OMICS DATA ANALYSIS

The large-scale and complicated omics data in combination with different clinical characteristics (Fig. 3 and 4) must be analyzed with appropriate statistical methods to reveal important signal molecule pattern [22] , signal pathways, molecular networks, and pathway-network systems [23] that operate in a specific biological system [2, 7] . The choice of an appropriate statistical method depends on the data characteristics and research goal. However, one must realize statistical significance and biological significance in those statistical and biological analysis processes. Statistical significance and biological significance are two different concepts with overlapping of their results. A statistically significant result must be reasonably interpreted with corresponding biological processes to decide its biological significance.

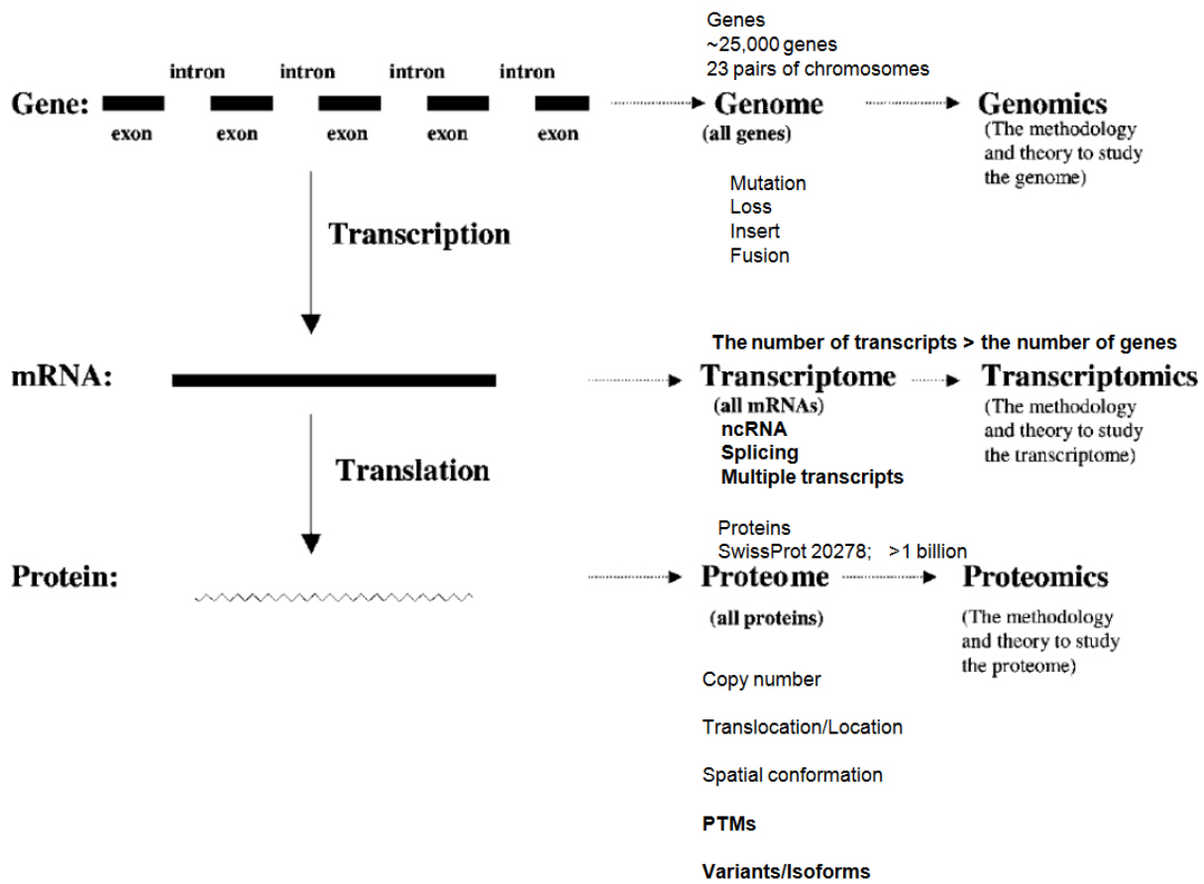

Fig. 4 Relationship among the genome, transcriptome, and proteome in cancer biology. r = 0.4 between transcriptome and proteome. Modified from Zhan and Desiderio (2005) [14] , with permission from Wiley-VCH, copyright 2005.

4 EXAMPLE TAKEN FOR STATISCAL CONSIDERATION AND BIOLGICAL SIGNIFICANCE

One example was taken below to clarify the statistical consideration and biological significance in the systems biology study. One study [7] was designed to discover statistically significant signaling pathways and networks with pituitary adenoma protein-mapping data [24] , comparative proteomic data [25, 26] , and nitroproteomic data [27-29] from Ingenuity Pathway Analysis (IPA) knowledge base (IPAKB) that contains a large-scale scientific findings and many canonical pathways and networks [7] . This study employed the Fisher's exact test in the IPA program to identify statistically significant pathways or networks with a significance level of 0.05. For this statistical analysis, the null hypothesis (H0) is that proteomic dataset of pituitary adenoma is not associated with all pathway networks that are stored in the IPAKB, the alternative hypothesis (H1) is that the proteomic dataset of pituitary adenoma is associated with all pathway networks in the IPAKB, and the level of significance is set as 0.05. The Fisher's exact test was used to calculate the *p*-value to determine the probability that the association between the molecules in the proteomic dataset and the canonical pathway

networks that is explained only by chance. The statistically significant result was that when $p < 0.05$, H0 was rejected, and H1 was accepted, which means that the identified signaling pathway networks exist in the pituitary adenoma. Based on this statistical hypothesis, method, and criteria of statistical significance, this study identified 12 significant canonical pathways (Fig. 5) and 1 network from qualitative nitroproteomic dataset in pituitary adenomas, 12 significant canonical pathways (Fig. 6) and 1 network from qualitative nitroproteomic dataset in controls, 9 significant canonical pathways (Fig. 7) and 3 networks from comparative proteomic dataset, and 37 significant canonical pathways (Fig. 8) and 6 networks from protein-mapping dataset.

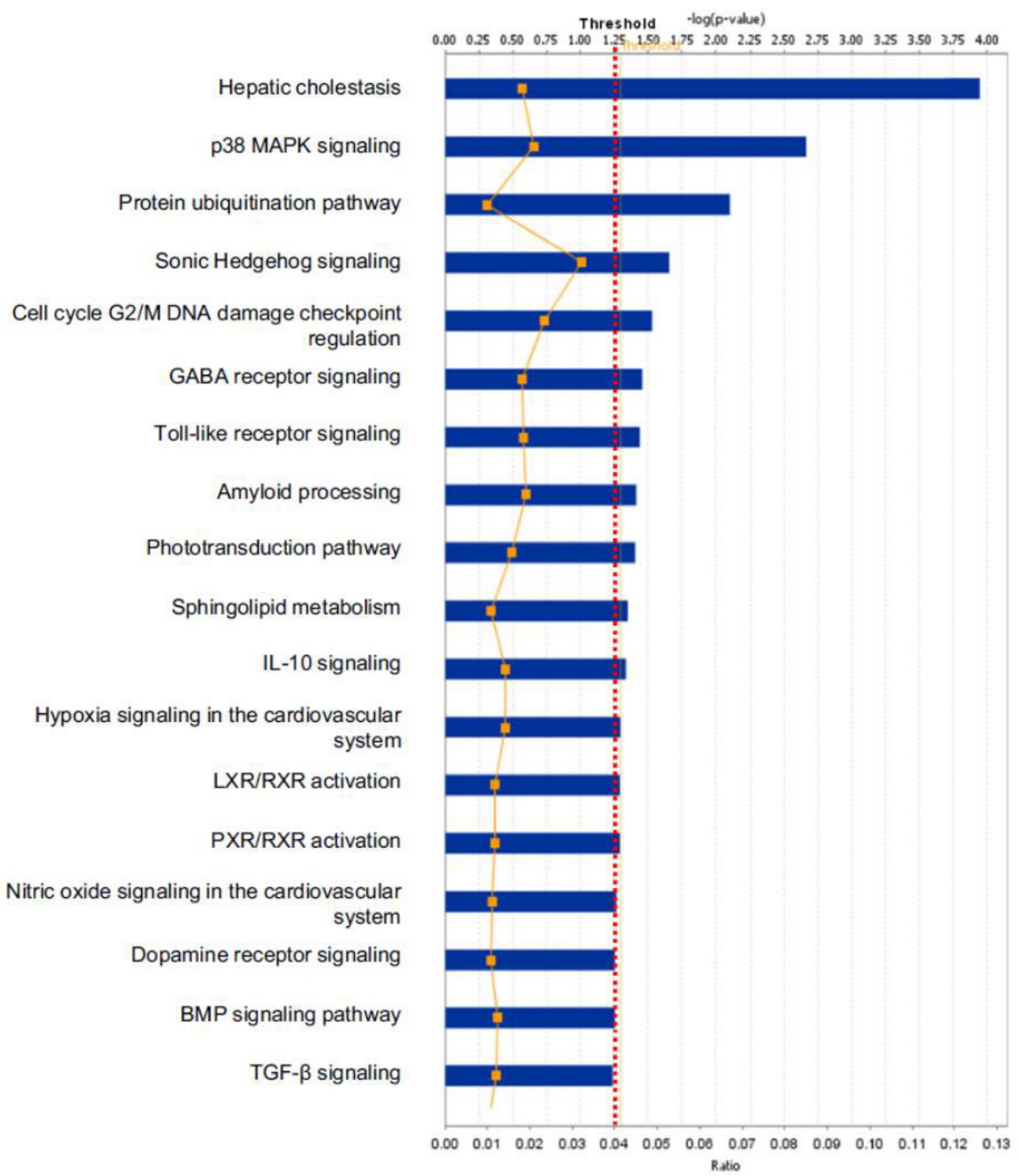

Fig. 5 Significant canonical pathways that are involved with pituitary adenoma nitroproteins. Reproduced from Zhan and Desiderio (2010) [7], with permission from BioMed Central Publisher. Copyright 2010 remains with authors due to the open-access article under the Creative Commons Attribution License.

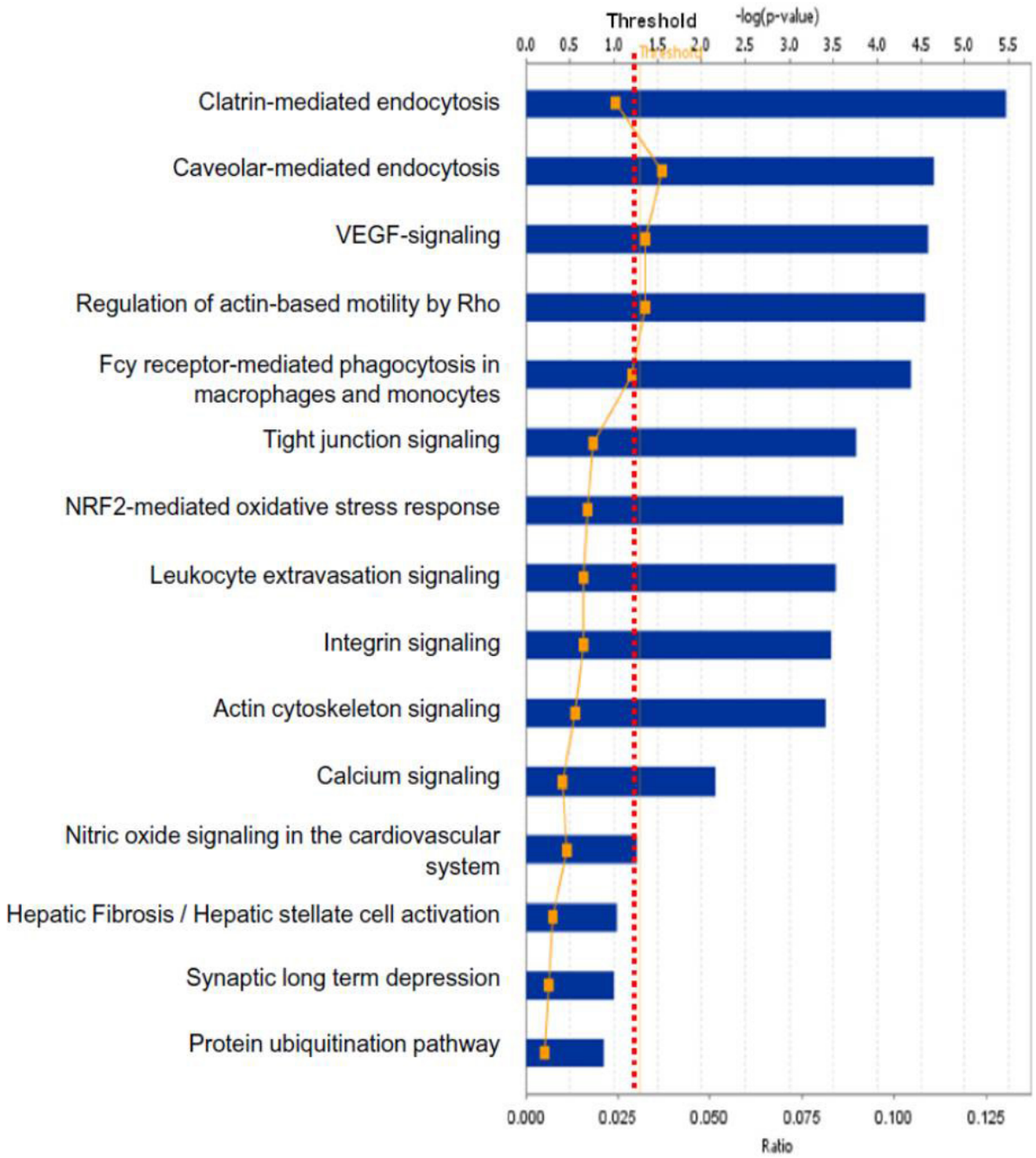

Fig. 6 Significant canonical pathways that are involved with control pituitary nitroproteins. Reproduced from Zhan and Desiderio (2010) [7] , with permission from BioMed Central Publisher. Copyright 2010 remains with authors due to the open-access article under the Creative Commons Attribution License.

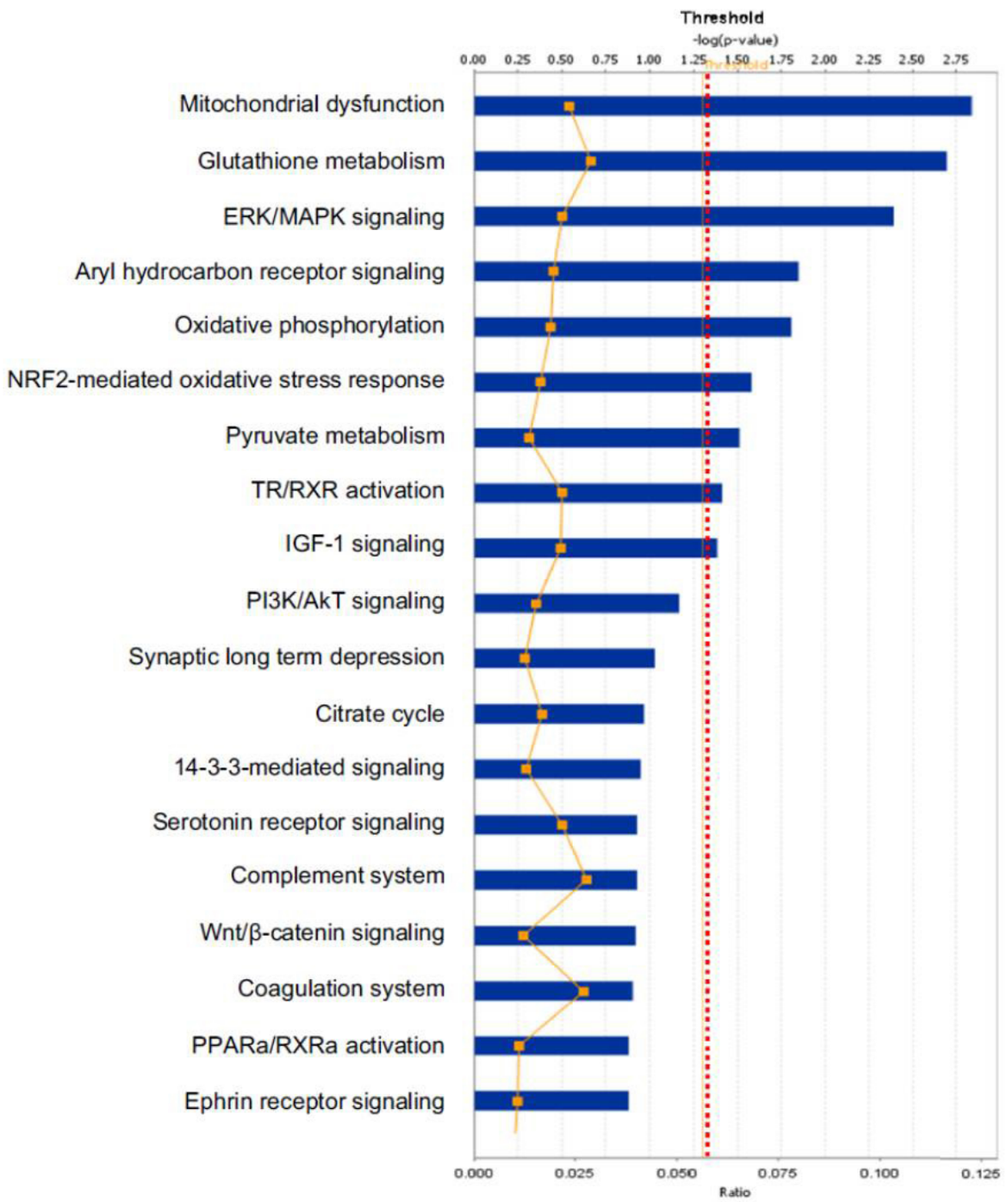

Fig. 7 Significant canonical pathways that are involved with pituitary adenoma comparative-proteomic data. Reproduced from Zhan and Desiderio (2010) [7] , with permission from BioMed Central Publisher. Copyright 2010 remains with authors due to the open-access article under the Creative Commons Attribution License.

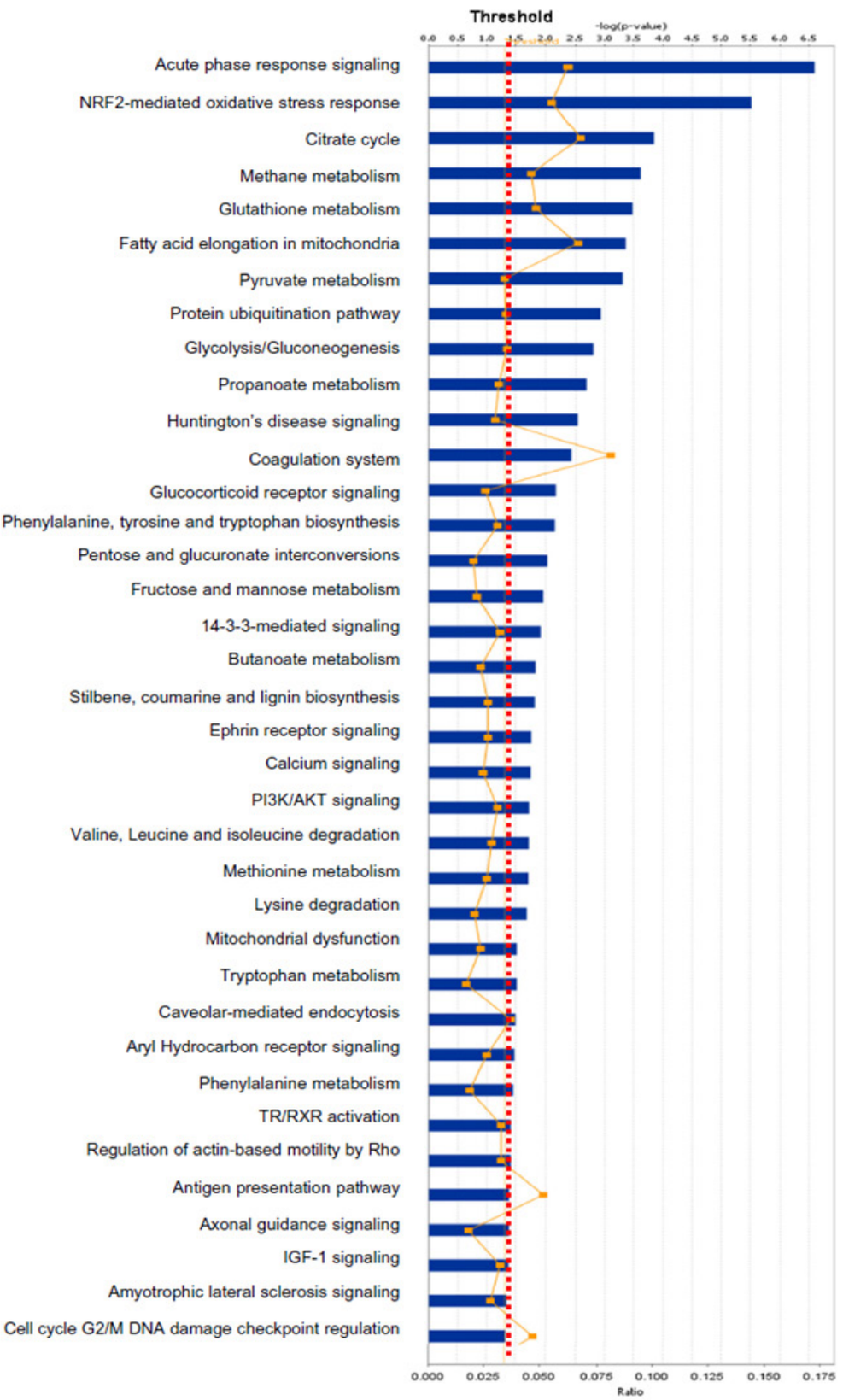

Fig. 8 Significant canonical pathways that are involved with pituitary adenoma protein-mapping data. Reproduced from Zhan and Desiderio (2010)[7], with permission from BioMed Central Publisher. Copyright 2010 remains with authors due to the open-access article under the Creative Commons Attribution License.

In this study, one can clearly find that if the significance level of 0.01 or 0.001 is used, it is more stringent criteria for this type of data analysis. It could decrease the probability of false positives, but it also leads to the loss of some biologically meaningful information. For example, if the significance level of 0.001 [or -log (0.001) = 3] is utilized, then there will be only 1 left significant canonical pathway (Fig. 5) identified from pituitary adenoma nitroproteomic data, 10 significant canonical pathways (Fig. 6) from control pituitary nitroproteomic data, no significant canonical pathways (Fig. 7) derived from comparative proteomics data, and 7 statistically significant canonical pathways (Fig. 8) from the protein-mapping data. As a result, compared to the level of statistical significance $p < 0.05$, many important canonical pathways (Fig. 5-8) are missing at the level of statistical significance $p < 0.01$ or 0.001. In fact, many differentially expressed proteins (DEPs) with a biological significance [25, 26] are derived from those important canonical pathways that were missed at the level of statistical significance $p < 0.01$ or 0.001; and here, the biological significance means the real variations or effects in a biological system. Furthermore, in Fig. 5, more stringent criteria ($p < 0.001$) would simply result in a significant canonical pathway – hepatic cholestasis. However, this pathway does not have much biological meaning for pituitary adenomas. On the other hand, the canonical pathways p38 MAPK signaling, cell-cycle G2/M DNA damage-checkpoint regulation, and protein-ubiquitination pathways (Fig. 5) were recognized as statistical significance with a significance level of 0.05, which can be reasonably linked to the real pituitary adenoma biological processes.

Any statistically significant result only serves as a reference for biological significance, and must be rationally interpreted with corresponding biological processes to decide its biological significance. Statistical significance does not mean a real variation or effect in a biological system. Some statistically significant results do not have any real biological meaning at al. A typical example is that hemoglobin is often identified as statistically significant DEP between pituitary adenoma and control tissues [25, 26] . However, it cannot be taken as a real DEP or biologically meaningful biomarker for a pituitary adenoma because its statistical significance is probably resulted from residual blood contamination. Another example, the canonical pathway hepatic cholestasis (Fig. 5) is ranked top one with a statistical significance, but it does not have any biological effect for pituitary adenoma pathogenesis. Moreover, for some cases, even though there might not be any statistical significance, those molecules still have biological significance. For instance, some genes have only a small change at the mRNA level without any significant difference; however, that small change at the mRNA level could result in an amplified change on the protein level, this result is still an interesting and meaningful finding. Therefore, when a research uses a certain statistical standard and receives corresponding statistical results, one should determine carefully whether the results are biologically relevant or just occur only by chance.

Based on these statistical considerations, those statistically significant pathways and networks identified with the Fisher's exact test with a significance level of 0.05 were reasonably explained within the pituitary adenoma biological system. Four important biological pathways [7] were identified for pituitary adenomas according to these considerations, including mitochondrial dysfunction, oxidative stress, cell-cycle dysregulation, and the MAPK signaling abnormality. These four biological pathways provide important clues and direction for further in-depth studies of pituitary adenomas.

5 MOLECULAR NETWORK CONCEPT-BASED STATISTICAL CONSIDERATION AND BIOLOGICAL SIGNIFICANCE

Molecules from genome, transcriptome, proteome, metabolome interact mutually to form an interactome to exert their biological functions in a biological system [30-32] , and this interactome is an ideal concept. Molecule networks are an important bridge to reach that interactome. Generally, all molecules are interacted and regulated mutually in the molecule-network system. Relative to the normal status, the altered molecule-network system occurs in a certain condition such as a cancer biological system. Based on the concept of molecular network, several issues are worth considering for statistical vs. biological significances: (i) Hub molecules would play very important roles in the molecule-network system; however, the amount of some hub-molecules would not alter much, and even no significant change in a disease status relative to the normal status. Thus, that hub molecule would do not have statistically significant alteration but have important biological function. A study found that hub-molecules and bottleneck-molecules in molecular networks were changed significantly slower than non-hub and non-bottleneck molecules, that the variation rate of hub-molecules was significantly lower than that of bottleneck-molecules, and that hub-molecules received stronger constraint than

bottleneck-molecules [33-35] . (ii) Each comparative omics analysis is commonly to identify differentially expressed genes (DEGs) and proteins (DEPs) with a certain changed fold with a statistical significance [25,26] . In fact, all molecules with statistical significance and without statistical significance interact mutually in a molecular network system, those molecules without statistical significance might still important in the molecular network system, and furthermore, the level of statistical significance (such as $p < 0.05$, $p < 0.01$, or $p < 0.001$) is determined by researchers during experimental design and before experiments. Thus, many biological information would be lost if one completely relies on statistical significant results. (iii) All molecules with or without differentially expressed molecules are used to establish the corresponding molecule networks[32, 36] that will be more representative result between two given conditions compared to that molecular networks derived from only statistically significant changed molecules. A human interactome [36] was constructed to connect 5400 proteins with 28,500 interactions in three quantitative dimensions using stoichiometries and abundances, which revealed that weak interactions dominate the network and have critical topological properties, and that there are rare stable complexes that stand out by a signature of balanced stoichiometries.

6 CONCLUSION

It is very complicated and important for statistical consideration and biological significance in a given big data and biological system. One must realize the difference and relationship between statistical and biological significances. The right statistical method must be selected for a given big data. The statistical results must be reasonably explained in a specific biological system. One must not use statistical significance to kidnap biological significance. Statistical significance is not equal to biological significance. Statistical result is only a reference to determine a biological significance.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 81572278 and 81272798 to XZ), the grants from China “863” Plan Project (Grant No. 2014AA020610-1 to XZ), the Xiangya Hospital Funds for Talent Introduction (to XZ), and the Hunan Provincial Natural Science Foundation of China (Grant No. 14JJ7008 to XZ).

AUTHOR’S CONTRIBUTION

X.Z. conceived the concept, collected pertinent references, designed and wrote the manuscript, and trained Y.L, X.H.Z, and Y.M regarding statistical significance, biological significance, and systems biology. Y.L. and Y.M participated in the collection of references, discussion and modification of manuscript. X.H.Z participated in revision of the manuscript. All authors approved the final manuscript.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests regarding the publication of this article.

REFERENCES

1. Grech G, Zhan X, Yoo BC, Bubnov R, Hagan S, Danesi R, Vittadini G, Desiderio D. EPMA Position Paper in Cancer: Current overview and future perspectives. EPMA J. 2015; 6: 9.
2. Hu R, Wang X, Zhan X. Multi-parameter systematic strategies for predictive, preventive and personalised medicine in cancer. EPMA J. 2013; 4: 2.
3. Hood L, Heath JR, Phelps ME, Lin B. Systems biology and new technologies enable predictive and preventative medicine. Science. 2004; 306 (5696): 640-643.
4. Hood L, Tian Q. Systems approaches to biology and disease enable translational systems medicine. Genomics Proteomics Bioinformatics. 2012; 10 (4): 181-185.
5. Tian Q1, Price ND, Hood L. Systems cancer medicine: towards realization of predictive, preventive, personalized and participatory (P4) medicine. J Intern Med. 2012; 271 (2): 111-121.
6. Golubnjitschaja O, Costigliola V, EPMA. General report & recommendations in predictive, preventive and personalized medicine 2012: White Paper of the European Association for Predictive, Preventive and Personalised Medicine. EPMA J. 2012; 3: 14.
7. Zhan X, Desiderio DM. Signal pathway networks mined from human pituitary adenoma proteomics data. BMC Med Genomics. 2010; 3: 13.

8. Altmae S, Esteban FJ, Stavreus-Evers A, Simon C, Giudice L, Lessey BA, Horcajadas JA, Macklon NS, D'Hooghe T, Campoy C, Fauser BC, Salamonsen LA, Salumets A. Guidelines for the design, analysis and interpretation of 'omics' data: focus on human endometrium. Hum Reprod Update. 2014; 20: 12-28.

9. Li Y, Wu FX, Ngom A. A review on machine learning principles for multi-view biological data integration. Brief Bioinform. 2016; 2016: 1-16.

10. Zhan X, Desiderio DM. The use of variations in proteomes to predict, prevent, personalize treatment for clinically non-functional pituitary adenomas. EPMA J. 2010; 1: 439-459.

11. Zhan X, Hu R, Wang X. Multi-parameter systematic strategy opinion that predicts, prevents, and personalized treats a cancer. EPMA J. 2014; 5 (Suppl 1): A25.

12. Zhan X. Systematic strategy opinion for research and clinical practice of chronic diseases. Int J Chronic Dis Ther. 2015; 1 (3e): 1-2.

13. Ferrer-Alcón M, Arteta D, Guerrero MJ, Fernandez-Orth D, Simón L, Martinez A. The use of gene array technology and proteomics in the search of new targets of diseases for therapeutics. Toxicol Lett. 2009; 186 (1): 45-51.

14. Zhan X, Desiderio DM. Comparative proteomics analysis of human pituitary adenomas: current status and future perspectives. Mass Spectrom Rev. 2005; 24: 783– 813.

15. Tyers M, Mann M. From genomics to proteomics. Nature. 2003; 422 (6928): 193-197.

16. Chung CH, Levy S, Chaurand P, Carbone DP. Genomics and proteomics: emerging technologies in clinical cancer research. Crit Rev Oncol Hematol. 2007; 61 (1): 1-25.

17. Haisch C. Raman-based microarray readout: a review. Anal Bioanal Chem. 2016; 408 (17): 4535-4545.

18. Bah A, Forman-Kay JD. Modulation of Intrinsically Disordered Protein Function by Post-translational Modifications. J Biol Chem. 2016; 291 (13): 6696-6705.

19. Zhan X. Insight into protein variants/isoforms and post-translational modifications in a proteome. A Proteomics. 2015; 2 (1): 1009.

20. Tebani A, Abily-Donval L, Afonso C, Marret S, Bekri S. Clinical Metabolomics: The New Metabolic Window for Inborn Errors of Metabolism Investigations in the Post-Genomic Era. Int J Mol Sci. 2016; 17: 1167.

21. Junot C, Fenaille F, Colsch B, Be´cher F. High resolution mass spectrometry based techniques at the crossroads of metabolic pathways. Mass Spectrom Rev. 2014; 33: 471–500.

22. Veenstra TD, Prieto DA, Conrads TP. Proteomic patterns for early cancer detection. Drug Discov Today. 2004; 9 (20): 889-897.

23. Zhang DY, Ye F, Gao L, Liu X, Zhao X, Che Y, Wang H, Wang L, Wu J, Song D, Liu W, Xu H, Jiang B, Zhang W, Wang J, Lee P. Proteomics, pathway array and signaling network-based medicine in cancer. Cell Div. 2009; 4: 20.

24. Zhan X, Desiderio DM: A reference map of a human pituitary adenoma proteome. Proteomics. 2003; 3 (5): 699-713.

25. Evans CO, Moreno CS, Zhan X, McCabe MT, Vertino PM, Desiderio DM, Oyesiku NM: Molecular pathogenesis of human prolactinomas identified by gene expression profiling, RT-qPCR, and proteomic analyses. Pituitary. 2008; 11 (3): 231-245.

26. Moreno CS, Evans CO, Zhan X, Okor M, Desiderio DM, Oyesiku NM: Novel molecular signaling and classification of human clinically nonfunctional pituitary adenomas identified by gene expression profiling and proteomic analyses. Cancer Res. 2005; 65 (22): 10214-10222.

27. Zhan X, Desiderio DM: The human pituitary nitroproteome: detection of nitrotyrosylproteins with two-dimensional Western blotting, and amino acid sequence determination with mass spectrometry. Biochem Biophys Res Commun. 2004; 325 (4): 1180-1186.

28. Zhan X, Desiderio DM: Nitroproteins from a human pituitary adenoma tissue discovered with a nitrotyrosine affinity column and tandem mass spectrometry. Anal Biochem. 2006; 354 (2): 279-289.

29. Zhan X, Desiderio DM: Linear ion-trap mass spectrometric characterization of human pituitary nitrotyrosine-containing proteins. Int J Mass Spectrom. 2007; 259 (1-3): 96-104.

30. Zhan X, Wang X, Long Y, Desiderio DM. Heterogeneity analysis of the proteomes in clinically nonfunctional pituitary adenomas. BMC Med Genomics. 2014; 7: 69.

31. Zhan X, Long Y. Exploration of molecular network variations in different subtypes of human nonfunctional pituitary adenomas. Front Endocrinol. 2016; 7: 13.

32. Cho S, Park SG, Lee DH, Park BC. Protein-protein interaction networks: from interactions to networks. J Biochem Mol Biol. 2004; 37 (1): 45-52.

33. Pang E, Hao Y, Sun Y, Lin K. Differential variation patterns between hubs and bottlenecks in human protein-protein interaction networks. BMC Evol Biol. 2016; 16(1): 260.

34. Kiran M, Nagarajaram HA. Interaction and localization diversities of global and local hubs in human protein-protein interaction networks. Mol Biosyst. 2016; 12(9): 2875-2882.

35. Ota M, Gonja H, Koike R, Fukuchi S. Multiple-Localization and Hub Proteins. PLoS One. 2016; 11(6): e0156455.

36. Hein MY, Hubner NC, Poser I, Cox JR, Nagaraj N, Toyoda Y, Gak IA, Weisswange I, Mansfeld J, Buchholz F, Hyman AA, Mann M. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell. 2015; 163: 712–723.