Editorial

Using gene set signature to guide hypothesis-driven experiments

Liang-Chuan Lai

Graduate Institute of Physiology, National Taiwan University, Taipei, Taiwan

Corresponding to: Liang-Chuan Lai. Graduate Institute of Physiology, National Taiwan University, Taipei 110, Taiwan. Email: llai@ntu.edu.tw.

Submitted Jan 20, 2013. Accepted for publication Feb 20, 2013.
doi: 10.3978/j.issn.2218-676X.2013.02.04
Scan to your mobile device or view this article at: http://www.thetcr.org/article/view/957/html

Complex diseases, such as cancer, are inherently multi-genetic mutations. Since it may be caused by a combinatorial effect of many mutations, the individual effect of each mutation may be too small to discover. In addition, the progresses of tumorigenesis not only include the mutations that facilitate tumorigenesis (called tumor drivers), but also include those accumulated during the growth of the tumor (known as tumor passengers) (1). Therefore, high degree of morphologic and clinical diversities exists among various types of cancers; intrinsic heterogeneity of cancer is very common among patients due to different genetic and environmental perturbations. However, traditional approaches of clinical disease classification are mainly based on pathological analysis of patients and existing knowledge of diseases. The current pathologic classification and ability to predict postsurgical prognosis are quite inadequate. Hence, although the existence of marked heterogeneity is well appreciated, virtually all cancers currently are treated similarly.

Recently, as advent in the post-genome era, a number of individualized molecular signature-based predictions that partition patients into distinct prognostic groups have been developed. This approach consists of examining a cohort of patients with a particular type of cancer, and is used to identify biomarkers. The biomarkers are usually used for classifying cancer subtypes with clinical or therapeutic implications. For disease classification, it adopts a supervised approach. Briefly, it starts with a set of samples with a known partition into disease subtypes (e.g., metastatic or not). The primary interest is to identify differentially expressed genes via quantitative statistical analysis to evaluate statistical significance of individual gene between two conditions. Once these differentially expressed genes have been identified, a clustering analysis is performed to group genes with similar expression patterns across different experimental conditions, and the biological meanings of observed expression changes are inferred using gene ontology (GO) or biological pathway-based analysis. After that, among these differentially expressed genes, various algorithms are used to identify a classifying principle based on the specific molecular features. In addition to disease classification, these biomarkers can also help determine whether a given patient will respond to a particular medicine or is susceptible to certain tailored therapy. However, criticism is frequently raised regarding that the genes used in such prognosis prediction classifiers have minimal overlaps among different reports, and provide little biological insight into the underlying mechanisms.

Alternatively, the other approach is to monitor the average differential expression level of genes belonging to a given functional category. This method is called gene set-wise differential expression analysis method. Gene set approaches based on the idea that complex diseases can be better understood from the perspective of dys-regulated gene sets than at the individual gene level. It has been successfully detected subtle but set-wise coordinated expression changes that cannot be detected by individual gene tests. Also, utilizing pre-defined and well-established gene sets rather than finding or creating novel lists of genes provides straightforward biological interpretation (2). The first developed in this category is the Gene Set Enrichment Analysis (GSEA), which evaluates the significant association with phenotypic classes for each priori defined gene set (3).

Nevertheless, implementing this idea requires a way to score candidate gene sets. Various methods have been suggested for measuring the significance of the differential expression of genes in a gene set. For example, averaged Z-value of genes (4), Gene Set Z-score (5), Kolmogorov-Smirnov (KS) statistics (6,7), Gene Enrichment Ranking (8), and Pearson correlation (9) were used previously to score gene set. However, among these methods, Z-value and KS statistics are effective only when member genes have consistent directional changes in expression. Also, many methods only consider the expression pattern and regard every gene with equal weight in the scoring scheme. Therefore, in order to project gene expression levels of a set of genes to a

scalar score, Hsiao *et al*. presented a scoring method (10) in this issue, adapted from their previous work (11), called Signature-score (S-score), and applied it to cholangiocarcinoma (CAC), an aggressive hepatic cancer that arises from bile duct cells. In this improved scoring algorithm, S-score concurrently evaluates both up- and down-regulated components of a gene set signature through Z-value and a sign function, and adjusts member genes' weights by *P*-value of t statistic. For determining the threshold for active gene set, Hsiao *et al*. define the qualitative boundary at the 99% prediction interval. Lastly, hierarchical clustering is used to help identify close-related gene sets. They applied this improved scoring algorithm to CAC, and identified one cluster positively correlated with the cell cycle and another cluster inversely correlated with immune function.

In the post-genome era, the real challenge of biologists is how to gain insight from the massive data, and translate the inferred gene sets into research questions of interest and hypothesis that can be tested. Gene set-based methods have been applied successfully in many disease studies, thereby reducing complexity and facilitating the generation of testable hypotheses. After all, biological validation of bioinformatics was a far more important end goal. Although, Hsiao *et al*. presented in the case of CAC, S-score is expected to be effectively applied to other complex diseases in assigning functional roles to disease-associated gene signature sets and in identifying potential therapeutic targets.

Acknowledgments

Funding: None.

Footnote

Provenance and Peer Review: This article was commissioned by the editorial office, *Translational Cancer Research*. The article did not undergo external peer review.

Conflicts of Interest: The author has completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.3978/j.issn.2218-676X.2013.02.04). The author has no conflicts of interest to declare.

Ethical Statement: The author is accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Baudot A, Real FX, Izarzugaza JM, et al. From cancer genomes to cancer models: bridging the gaps. EMBO Rep 2009;10:359-66.
2. Kim JH. Chapter 8: biological knowledge assembly and interpretation. PLoS Comput Biol 2012;8:e1002858.
3. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545-50.
4. Ebi H, Tomida S, Takeuchi T, et al. Relationship of deregulated signaling converging onto mTOR with prognosis and classification of lung adenocarcinoma shown by two independent in silico analyses. Cancer Res 2009;69:4027-35.
5. Törönen P, Ojala PJ, Marttinen P, et al. Robust extraction of functional signals from gene set analysis using a generalized threshold free scoring function. BMC Bioinformatics 2009;10:307.
6. Barbie DA, Tamayo P, Boehm JS, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 2009;462:108-12.
7. Accetturo M, Creanza TM, Santoro C, et al. Finding new genes for non-syndromic hearing loss through an in silico prioritization study. PLos One 2010;5:e12742.
8. Luo B, Cheung HW, Subramanian A, et al. Highly parallel identification of essential genes in cancer cells. Proc Natl Acad Sci U S A 2008;105:20380-5.
9. Gibbons DL, Lin W, Creighton CJ, et al. Expression signatures of metastatic capacity in a genetic mouse model of lung adenocarcinoma. PLoS One 2009;4:e5401.
10. Hsiao TH, Chen HH, Lu JY, et al. Utilizing signature-score to identify oncogenic pathways of cholangiocarcinoma. Transl Cancer Res 2012. [Epub ahead of print].
11. Rubin BP, Nishijo K, Chen HI, et al. Evidence for an unanticipated relationship between undifferentiated pleomorphic sarcoma and embryonal rhabdomyosarcoma. Cancer Cell 2011;19:177-91.

Cite this article as: Lai LC. Using gene set signature to guide hypothesis-driven experiments. Transl Cancer Res 2013;2(1):1-2. doi: 10.3978/j.issn.2218-676X.2013.02.04