Genomics Proteomics Bioinformatics 12 (2014) 31–38

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Pathway-based Analysis of the Hidden Genetic Heterogeneities in Cancers

Xiaolei Zhao [1,#], Shouqiang Zhong [2,#], Xiaoyu Zuo [3], Meihua Lin [1], Jiheng Qin [1], Yizhao Luan [1], Naizun Zhang [2], Yan Liang [2,*], Shaoqi Rao [1,3,*]

[1] *Institute for Medical Systems Biology and Department of Medical Statistics and Epidemiology, School of Public Health, Guangdong Medical College, Dongguan 523808, China*
[2] *Maoming People's Hospital, Maoming 525000, China*
[3] *Department of Medical Statistics and Epidemiology, School of Public Health, Sun Yat-Sen University, Guangzhou 510080, China*

Received 10 October 2013; revised 6 December 2013; accepted 9 December 2013
Available online 22 January 2014

Handled by Arndt G. Benecke

KEYWORDS

Genetic heterogeneity;
Pathway-based approach;
Sample partitioning;
Enrichment analysis;
Survival analysis;
Cancer

Abstract Many cancers apparently showing similar phenotypes are actually distinct at the molecular level, leading to very different responses to the same treatment. It has been recently demonstrated that pathway-based approaches are robust and reliable for genetic analysis of cancers. Nevertheless, it remains unclear whether such function-based approaches are useful in deciphering molecular heterogeneities in cancers. Therefore, we aimed to test this possibility in the present study. First, we used a NCI60 dataset to validate the ability of pathways to correctly partition samples. Next, we applied the proposed method to identify the hidden subtypes in diffuse large B-cell lymphoma (DLBCL). Finally, the clinical significance of the identified subtypes was verified using survival analysis. For the NCI60 dataset, we achieved highly accurate partitions that best fit the clinical cancer phenotypes. Subsequently, for a DLBCL dataset, we identified three hidden subtypes that showed very different 10-year overall survival rates (90%, 46% and 20%) and were highly significantly ($P = 0.008$) correlated with the clinical survival rate. This study demonstrated that the pathway-based approach is promising for unveiling genetic heterogeneities in complex human diseases.

Introduction

Genetic heterogeneity has attracted increasing attention in the study of genetic mechanisms of complex diseases. It describes the biological complexities that apparently similar characters may result from different genes or different genetic mechanisms [1]. In the clinical setting, patients with diseases displaying a similar phenotype but resulting from different genetic causes frequently respond very differently to the same

* Corresponding authors.
E-mail: lye30668@aliyun.com (Liang Y), raoshaoq@gdmc.edu.cn (Rao S).
# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.gpb.2013.12.001

treatment and thus receive a markedly different prognosis. Therefore, elucidation of the genetic heterogeneities underlying complex diseases has profound influences on both modern clinical practice and basic biomedical research.

Rapidly accumulated genomic-scale molecular data provide good opportunities to unveil the genetic heterogeneities in complex diseases at the molecular level. Significant improvements in methods and applications for analysis of the genetic heterogeneity have been achieved in the past decades. The usefulness of large-scale gene expression data, as measured by microarrays, has noticeably been indicated by the successful stratification of diffuse large B-cell lymphoma (DLBCL) [2–5]. In these pioneering studies, an unsupervised clustering algorithm was used to partition both gene expression data and patients with an aim to define genetically homogeneous novel cancer subgroups among cancer patients based on the principle that patients within the same cluster probably involve the similar molecular pathogenesis and hence could be grouped into the same molecular subphenotype [6]. Although the traditional clustering analysis based on individual gene expression profiles has achieved great success in unveiling the genetic heterogeneity, it seldom considered the combined actions of multiple functionally dependent genes. It is increasingly recognized that complex diseases such as cancers are a consequence of alterations in a complicated cascade of events involving multiple biological processes and pathways. Thus, subtypes identified by individual genes often lack good biological interpretations. In this sense, the development of function-based methods for cancer subtyping is warranted.

Gene Ontology (GO) [7] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [8] are the two most common databases currently used for gene functional annotation. GO terms are used primarily for the annotation of individual gene products, whereas KEGG pathway terms are used for the annotation of classes of gene products, thus providing a more precise delineation of functionalities for a group of genes that act together to some extent. KEGG pathway is a collection of manually drawn pathway maps that represent the knowledge on molecular interactions and reaction networks for human diseases, environmental information processing, genetic information processing, *etc*., thus possibly providing biological interpretations of higher-level systemic functions [9]. Hence, a pathway-based approach can integrate the effects of genetic factors and biological networks [10] and has been used for disease classification [11]. In our previous work [1], we proposed a GO-based approach to unveil the hidden heterogeneities in cancers, and demonstrated that it can successfully integrate the cellular function and the gene expression profile, and the approach showed the greater advantage of GO in classifying the cancer types. In principle, a similar pathway-based approach should have comparable performance in the genetic analysis of molecular heterogeneities in cancers. Numerous studies have shown that the cancer subtypes are, in essence, related to multiple pathways [12–14]. For example, recent evidence has shown that molecular subtypes of DLBCL arise from distinct genetic pathways [15]. Therefore, this study aimed to verify whether a pathway-based approach is useful in deciphering molecular heterogeneities in complex diseases such as cancer.

In this study, we proposed a pathway-based clustering approach to unveil disease heterogeneities based on multiple pathways. First, we selected differentially expressed genes that are associated with specific disease conditions. It should be noted that algorithms such as the *t* test or *F* test are not proper for selecting differentially expressed genes due to the presence of genetic heterogeneity, because the validity of these tests relies on accurately and unambiguously defining phenotype characteristics. Hence, we took a robust metric, the overall variability of gene expression, to guide gene selection. Firstly, genes with top-ranked expression variations across samples, which explain most of the total variance potentially contributed by known or unknown factors (for example, the hidden cancer subtypes), were selected as ''feature genes'' in the initial gene selection as implemented in several previous studies [16,17]. Then, we identified KEGG pathways enriched with feature genes as ''putative signature pathways'' (here, ''enriched'' means that a pathway has saliently more feature genes (with large variance) than a random gene set of the same size does). Finally, we classified samples to identify the hidden disease subtypes using the expression profiles of genes annotated to these well-characterized pathways. In the numerical analysis, we first validated the proposed approach in accurately partitioning cancer phenotypes using a publicly-available large cancer dataset. Subsequently, we used the approach to identify the hidden subtypes of a notoriously heterogeneous phenotype, DLBCL. Our results demonstrated that three new subtypes identified using signature pathways had very different 10-year overall survival rates, and the partitions were highly significantly correlated with the clinical survival rates.

Results

Validation of the proposed pathway-based approach using a large microarray dataset

We selected the signature pathways that were significantly (FDR $\leqslant$ 0.01, see the Materials and methods section for the details) enriched with the 10% top-ranked genes with largest expression variances based on the NCI60 dataset [18]. As a result, three pathways were identified, which were used for the subsequent analyses. These include the small cell lung cancer pathway (hsa05222), the extracellular matrix (ECM)–receptor interaction pathway (hsa04512) and the focal adhesion pathway (hsa04510) (**Table 1**). First, we evaluated the ability of each signature pathway to accurately partition the samples into the known cancer types using the clustering analysis based on only the expression profiles of genes within the pathway. Our results based on each of the three pathways agreed well with the original clinical labels. The observed values for the adjusted Rand index (ARI) [19] (to measure the agreement between the identified clusters and the original partitions, ranging from 0 to 1, see the Materials and methods section for the details) were 0.83, 0.69 and 0.78, respectively. Subsequently, to determine the empirical significance of each pathway, we randomly selected 1000 gene subsets of the same pathway size from the null distribution as described in the Materials and methods section. No random subset achieved an ARI value higher than that of the corresponding pathway such that all identified signature pathways showed significantly better performance ($P < 0.001$) in correctly partitioning the samples (that is, more likely relevant to the phenotypic partitions). Furthermore, after applying the majority rule voting for integrating results from the three signature pathways, we