Chapter 31
Cross-Platform Microarray Data Integration Combining Meta-Analysis and Gene Set Enrichment Analysis

Jian Yu
Tongji University, China

Jun Wu
Shanghai Center for Bioinformation Technology, China

Miaoxin Li
Shanghai Center for Bioinformation Technology, China

Yajun Yi
Vanderbilt University, USA

Yu Shyr
Vanderbilt University, USA Yixue Li Shanghai Center for Bioinformation Technology, China

Lu Xie
Shanghai Center for Bioinformation Technology, China

ABSTRACT

Integrative analysis of microarray data has been proven as a more reliable approach to deciphering molecular mechanisms underlying biological studies. Traditional integration such as meta-analysis is usually gene-centered. Recently, gene set enrichment analysis (GSEA) has been widely applied to bring gene-level interpretation to pathway-level. GSEA is an algorithm focusing on whether an a priori defined set of genes shows statistically significant differences between two biological states. However, GSEA does not support integrating multiple microarray datasets generated from different studies. To overcome this, the improved version of GSEA, ASSESS, is more applicable, after necessary modifications. By making proper combined use of meta-analysis, GSEA, and modified ASSESS, this chapter reports two workflow pipelines to extract consistent expression pattern change at pathway-level, from multiple microarray datasets generated by the same or different microarray production platforms, respectively. Such strategies amplify the advantage and overcome the disadvantage than if using each method individually, and may achieve a more comprehensive interpretation towards a biological theme based on an increased sample size. With further network analysis, it may also allow an overview of cross-talking pathways based on statistical integration of multiple gene expression studies. A web server where one of the pipelines is implemented is available at: http://lifecenter.sgst.cn/mgsea//home.htm.

DOI: 10.4018/978-1-4666-3604-0.ch031

Copyright © 2013, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
INTRODUCTION AND BACKGROUND

Gene expression profiling has become an important tool for biological research, creating exponentially increasing amount of microarray data. However, several studies have shown that when identifying Differentially Expressed Genes (DEGs) in two groups of tissues (such as cancer vs. normal), the consistency of results among different labs or even batches was rather poor (Choi, 2004). This would cause confusion in follow-up experimental research. The direct comparison of various microarray studies is restricted by different protocols, microarray platforms, and analysis techniques, etc. (Warnat, 2005). Integrative analysis of multiple microarray data has always been both an attraction and a challenge.

There have been some studies using statistical and computational methods to integrate gene expression data (Warnat, 2005; Cheadle, 2007; Bosotti, 2007; Li 2006). The most typical integrative approach at gene level is meta-analysis. In statistics, a meta-analysis combines the results of several studies that address a set of related research hypotheses. In statistics, a meta-analysis combines the results of several studies that address a set of related research hypotheses. Meta-Analysis has been widely used in public health research (Berry, 2000). Rhodes (2002) and Choi (2004) were the first to apply meta-analysis in cross-platform microarray data integration. They identified the concordance of significantly differentially expressed genes among several microarray datasets of prostate cancer and hepatocellular carcinoma (HCC). It was demonstrated that meta-analysis increased the sensitivity of analysis and allowed for small but consistent expression changes to be detected. More applications of meta-analysis in gene expression studies were reported in the following studies (Hong, 2006; Conlon, 2007; Marot, 2009; Pihur, 2009).

Although meta-analysis succeeded in combining results from different datasets, it has its limitations. First limitation is, only overlapping genes represented in all platforms could be used in the analysis. Therefore it is more suitable for integration of studies using the same microarray platform. Same platforms indicate that the microarray chips are from the same manufacturer, same probe type (cDNA or oligo nucleotides) and same probe set design. Another limitation of meta-analysis is, it is an integrative analysis at single gene level, which means, results are often demonstrated as long lists of statistically significant genes without meaningful biological interpretations.

Recently, a gene set-level approach named Gene Set Enrichment Analysis (GSEA) has been widely applied in microarray study. It is a powerful technique to determine whether members of a predefined gene set, e.g., genes that belong to the same pathway or share the same cellular function or component are significantly changed in two groups of tissues based on the whole-genome gene expression data (Subramanian, 2005). For biological research, “pathway” gene sets are often chosen. Therefore, in this chapter, gene set-level is often referred to as pathway-level.

These existing studies suggest that for datasets from same microarray platform, meta-analysis plus GSEA may be a good approach for integrative analysis at pathway-level. In fact, that is our proposed Pipeline I in this chapter.

However, the situation may be more complex for integrative analysis across different microarray platforms. Platforms from the same manufacturer but with different probe sets or platforms from different manufacturers are defined as different platforms. For data generated using different platforms, a large amount of gene expression information would be lost if only genes present in all platforms are used, as required by meta-analysis. Therefore meta-analysis can not be directly performed. As for GSEA, one good example was shown by Subramanian (2005) who performed GSEA on two independent studies of gene expression profiling of lung adenocarcinoma patients with good or poor clinical outcome, and he identified overlap among the significantly enriched
Related Content

Informatics for Medicines Management Systems in Resource-Limited Settings
[www.igi-global.com/chapter/informatics-medicines-management-systems-resource/64063?camid=4v1a](www.igi-global.com/chapter/informatics-medicines-management-systems-resource/64063?camid=4v1a)

Computer-Aided Drug Design and Biological Evaluation of Novel Anti-Greek Goat Encephalitis Agents

Connecting Microbial Population Genetics with Microbial Pathogenesis: Engineering Microfluidic Cell Arrays for High-Throughput Interrogation of Host-Pathogen Interaction
[www.igi-global.com/chapter/connecting-microbial-population-genetics-microbial/76092?camid=4v1a](www.igi-global.com/chapter/connecting-microbial-population-genetics-microbial/76092?camid=4v1a)

Ubiquitous Healthcare
Theodor Panagiotakopoulos, Maria-Anna Fengou, Dimitrios Lymberopoulos and Eduard Babulak (2010). *Biocomputation and Biomedical Informatics: Case Studies and Applications* (pp. 254-280).
[www.igi-global.com/chapter/ubiquitous-healthcare/39618?camid=4v1a](www.igi-global.com/chapter/ubiquitous-healthcare/39618?camid=4v1a)