Integrative Factor Analysis – an unsupervised method for quantifying cross-study consistency of gene expression data

Xin Victoria Wang, Giovanni Parmigiani

PII: S0888-7543(17)30076-9
DOI: doi: 10.1016/j.ygeno.2017.08.009
Reference: YGENO 8916

To appear in: *Genomics*

Received date: 15 June 2017
Revised date: 28 August 2017
Accepted date: 30 August 2017

Please cite this article as: Xin Victoria Wang, Giovanni Parmigiani, Integrative Factor Analysis – an unsupervised method for quantifying cross-study consistency of gene expression data, *Genomics* (2017), doi: 10.1016/j.ygeno.2017.08.009

Integrative Factor Analysis – an unsupervised method for quantifying cross-study consistency of gene expression data

XIN VICTORIA WANG *, GIOVANNI PARMIGIANI

Department of Biostatistics and Computational Biology,
Dana-Farber Cancer Institute,
Boston, MA 02215, USA
Department of Biostatistics,
Harvard School of Public Health,
Boston, MA 02115, USA
vwang@jimmy.harvard.edu

Abstract

Integrative analyses of multiple gene expression studies are frequently performed. In the setting of two studies, integrative correlation (IGC) can be used to assess the consistency of co-expression of a given gene. For three or more studies, an extension of IGC gives a global score per gene. We propose to extend IGC and use factor analysis to assess the study-specific consistency of co-expression of genes when there are three or more studies, possibly on different platforms. Our method is able to identify studies whose expression patterns are different from others. Filtering genes based on our score is shown to improve the concordance of association with phenotype across studies.

1 Introduction

Thousands of microarray and RNA-seq data sets have been produced for the study of gene expression in many different diseases and phenotypes. Because of limitations in cost and access to appropriate samples, many of these data sets have a small number of samples. However, multiple data sets from different studies are often available to address the same or similar biological questions. There is enormous interest in utilizing these data sets in an integrated way so that better biological insights can be provided. But because there is usually a large variation between studies due to differences in platform, sample quality, reagent, and many other factors that can influence the measurements of gene expression, integrating microarray data across studies has not been a trivial task.

While the method for integrating studies is important, choosing appropriate genes and studies to include in the integrated analysis is equally critical. Not only do studies vary in terms of technology and sample

*To whom correspondence should be addressed.