International Journal of Biomedical Data Mining

ISSN: 2090-4924

Spampinato and Cavallaro, Int J Biomed Data Min
2016, 5:1
DOI: 10.4172/2090-4924.1000117

Short Communication Open Access

Meta-Analysis of Genomic Data: Between Strengths, Weaknesses and New Perspective

Spampinato AG and Cavallaro S*

Institute of Neurological Sciences, Italian National Research Council, Catania, Italy

***Corresponding author:** Sebastiano Cavallaro, Institute of Neurological Sciences, Italian National Research Council, Catania, Italy; Tel: +390957338111; E-mail: sebastiano.cavallaro@cnr.it

Rec date: Dec 22, 2015; **Acc date:** Jan 19, 2016; **Pub date:** Jan 29, 2016

Abstract

The rapid advances in high-throughput technologies, such as microarrays have revolutionizing the knowledge and understanding of biological systems and genetic signatures of human diseases. This has led to the generation and accumulation of a large amount of genomic data that need to be adequately integrated to obtain more reliable and valid results than those from individual experiments. Meta-analysis of microarray data is one of the most common statistical techniques used for combining multiple data sets. Despite its remarkable successes in discovering molecular subtypes, underlying pathways and biomarkers for the pathological process of interest, this method possesses several limitations.

Here, we provided a briefly overview of current meta-analytic approaches together with the basic critical issues in performing meta-analysis of genomic data, with the aim of helping researchers to evaluate the quality of existing, published data and obtain more detailed information on what will be the best strategy to adopt to execute a good meta-analysis.

Introduction

In the last decades, the rapid advances in high-throughput technologies, including sequencing and microarray assay, have transformed biomedical research by allowing comprehensive monitoring and deciphering of biological systems and the genome-wide discovery of diagnostic and prognostic gene signatures of human diseases. This has led to the generation and accumulation of significant amounts of high-throughput genomic data, many of which are deposited in several large publically-available data repositories, such as Gene Expression Omnibus (GEO) and Array Express [1,2].

The existence of such a large amount of existing genomic data has required the development of more robust and efficient statistical and computational resources and expertise than low-throughput technologies, which allow to retrieve, filtering, integrate, and compare this multitude of "-omics" data from different, independent studies cohesively into a single analysis [3,4]. In this regard, meta-analysis of microarray data represents one of the most common statistical techniques used for combining multiple data sets, offering considerable advantages in both overcoming individual study-specific biases and increasing statistical power to obtain more reliable and more valid results than those from individual experiments. Meta-analysis methods have already demonstrated to be very useful tools in bio-medical research, enabling researchers to discover disease subtypes, new biomarkers and therapeutic targets and biological pathways associated with the process of interest [5,6]. Despite those successes, current approaches for meta-analysis possess several limitations.

In the present paper, we provided a brief overview of current meta-analytic approaches and basic critical issues in performing meta-analysis of microarray data, with the aim of helping researchers to evaluate the quality of existing, published data and obtain more detailed information on what will be the best strategy to adopt to execute a good meta-analysis.

Overview of the main critical issues

The general workflow for the meta-analysis approach consists of three main steps: data selection, data preparation, including the processing of gene expression datasets and probe annotations, and data analysis. The quality of meta-analysis depends upon the quality of each individual microarray dataset as well as standardized procedures, algorithms for cluster analysis and methods utilized for the analysis. Considering that, the phase of microarray data processing provides for the transformation of image data (i.e. spot brightness) to gene expression values, it appears evident that one of the first problems encountered in performing a meta-analysis is the presence of a strong background and noise signal value (e.g., due to unspecific hybridization or spatial artifacts) that may affect data quality leading to a uncorrected and more random distribution of probe signals among probe sets. The development of methods and procedures aimed to quantify and improve the 'signal-to-noise ratio' are, therefore, necessary for conducting a good meta-analysis.

In addition to technical issues, it should be considered that one of the main goal in performing meta-analysis of microarray data is to obtain a more or less restricted set of genes whose expression is associated with the variable of interest, as in the case of pathological versus healthy conditions. Many current meta-analytic approaches select genes on the basis of univariate summary statistic parameters, including P-value. Applying these methodologies would require that all of the studies included in the analysis originate from relatively homogeneous sources, test the same hypotheses and/or are carried out under comparable conditions or treatments. However, biological, experimental and technical variations that occur between different

studies of the same phenotype/phenomena create substantial differences in results, obstructing the achievement of a good accuracy and reproducibility of the data derived from microarray experiments. In addition to, it is also important to note that, in the majority of cases, microarray data included in the meta-analysis are produced by using different platforms (e.g., Agilent, Affymetrix, Illumina) and, therefore, subjected to different processes of normalization and data analysis, making more difficult interpreting results in the context of other microarray experiments. For these reasons, in order to reduce the variability due to different pre-processing algorithms, it is often necessary to have access to the raw data of a gene expression experiment during the phase of data selection of meta-analysis. Once raw data have been obtained, in fact, they may be converted to information by using homogeneous assembly and normalization processes, allowing performing subsequent optimal downstream statistical and bio-informatics analyses of genomic data.

Raw gene expression datasets are generally obtained both as available data from different laboratories and/or from a systematic search in the online databases, including Gene Expression Omnibus (GEO) or Array Express. To uniform procedures of microarray raw data and to obtain, description, submission and, consequently, facilitate reproducible researches, guidelines have been established and included in the Minimum Information about Microarray Experiment (MIAME) standard functions from the MGED (Microarray Gene Expression Data) Society [7]. Moreover, an additional extension of MIAME concepts has been developed, The Standard Micro Array Reporting Template (SMART), which allows a specific gene list to be adequately recorded and described, making data accessible, comparable, and dynamically updatable. Unfortunately, despite the development and the adoption of these guidelines, the majority of studies submitted in public repositories are not MIAME compliant and the raw data are not always available, leading to insufficient annotations of experimental and bio-informatics approaches that is the main cause for the lack of reproducible research [8]. Therefore, the use of corrected and standardized data format, storage and quality remains, to date, a major challenge for the future of meta-analysis.

Independently of the biological, technical, and analytical procedures, microarray studies cannot be effectively compared without the use of opportune software that translates probe DNA sequences into biological meaning. In fact, another important pre-processing step in meta-analysis comparison is correlating probes to their corresponding genes within and across the different microarray platforms. This procedure allows often performing meta-analysis by selecting and integrating only genes that are present across the different platforms and removing those absent in one or more platforms, decreasing the number of genes with a consequent loss of information potentially important in the understanding of the phenomenon under investigation.

Generally, the relationship between probes and genes is determined by using the annotation information included in several public repositories, such as Genbank, UniGene and Entrez Gene identifiers, or other additional molecular biology databases. Although annotations and biological information are stored in relational databases, they are in many case distributed and shared as text files and included in the flat file databases. This data organization does not allow seamlessly integrating the heterogeneous sources of genomic data information, hampering the development of a simple and robust solution for an accurate and high-volume comparison of different gene expression profiles. This 'linguistic' disparity is even more evident when a cross-species comparative meta-analysis is performed. Besides to the cross-platform mapping of probes, in this case there is also the difficulty of comparing data between different organisms. Indeed, due to the complexity of evolutionary changes, such as gene duplication, there is not a correspondence across genes from different species, making difficult the comparison between their expression profiles. This comparison is often performed by using different databases including cluster of genes homologous/orthologous, like Homologue, through which it is possible to find homologous genes among those annotated of several completely sequenced eukaryotic genomes. Although these publicly available resources represent important tools to compare microarray studies between different organisms, it should be kept in mind that evolutionary orthology does not necessarily have a strong correspondence to function similarity and, thus, genetic and genomic alterations occurring in animal models of a particular human condition, not necessarily will have a similar impact on gene expression in humans.

New approaches and recent developments in meta-analysis methods

Current methods for performing meta-analysis present several limitations and the integration of microarray datasets from different studies still represents a significant computation and technical challenge. Collecting data from microarray repositories, identifying of available studies with consistent information, raw data re-processing and analysis and low-quality datasets represent, as previously said, concrete issues reducing meta-analysis efficiency. To overcome these limitations, new approaches have been proposed to conduct meta-analysis of gene expression data.

The development, for example, of Microarray meta-analysis database (M2DB) promises to improve the comparability between human microarray data by using a uniformly raw pre-processing, high-quality controlled data and microarray annotations manually curated through controlled vocabularies, based on information derived from scientific publications and online databases, like GEO and Array Express. More uniform data preprocessing allows eliminating the variance that occurs during microarray data transformation, improving, among other things, background correction, probe-set summarization and data normalization.

Among methods developed to increase the power of "ordinary" meta-analysis, particular attentions should be given to the Bi-level approach, the Elastic Net, SMA (Sequential Meta-analysis) and web-based tool Network Analyst [9-12].

The Bi-level approach constitutes a novel method to performing meta-analysis of gene expression data, analyzing these in a context of known biological pathways [11]. This method permits to integrate multiple independent studies for the same disease by performing an analysis on two levels. In the "intra-experiment", the analysis consists in splitting dataset into m smaller datasets. To this end, the statistical tests are performed independently and then p-values obtained are combined each other. In second level, an "inter-experiment" analysis is conducted, in which the algorithm conducts a statistical test for each individual experiment and then combines processed p-values. P-value calculation is done, for each of the m datasets, by using a pathway analysis method for each one of k pathways included in existing pathway databases. This technique has been demonstrated to improve the power of meta-analysis thanks to the bi-level framework that confer more robustness against bias, minor sensitive to outliers than

other traditional methods, and greater sensitivity in detecting small signal changes.

An additional methodological approach for conducting a powerful meta-analysis uses the versatile method of the Elastic Net for classification and regression. In statistics as well as in the fitting of linear or logistic regression model, Elastic Net functions as a regularized regression method that linearly combines the penalties function of LASSO (least absolute shrinkage and selection operator) and Ridge methods. In this framework [12] Elastic Net permits to build a predictive model based on gene expression data and other variables, such as patient characteristics. In particular, Elastic Net analyzes the merged data into a single matrix deriving from a cross-study normalization procedure of raw data and is able to handle both continuous and categorical features. Moreover, through the application of a predictive model and the quantitative and qualitative determination of genes belonging to the 'expression signature' of the conditions of interest, the Elastic Net function permits to incorporate additional variables to the gene expression profiles, revealing the correlation between gene expression and corresponding covariates and is not strictly correlated with the biological phenomenon of interest.

The Sequential Meta-Analysis (SMA) is an approach aimed to find significant gene expression signatures by merging multiple microarray studies in chronological order, avoiding type I errors. Whit regard to traditional meta-analysis methods, this approach could also represents a useful tool to evaluate if a greater number of experiments is needed to draw a conclusion. In fact, for each gene of interest, SMA assesses whether collected samples already show sufficient evidences for a certain effect size or if further experiments should be added [10].

Finally, the web-based tool Network Analyst has been developed to perform common and complex meta-analysis, with both advanced statistics and visualization strategies for allowing an efficient data comparison [9]. In light of the growing amount of publicly available gene expression data as well as the increasingly recourse to comparative analysis among different microarray studies and platforms, it appears evident the need to further develop new methodologies and refine existing methods to improve data quality, eliminate platform-specific bias and permit better cross-platform normalization processes and statistical analyses.

References

1. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, et al. (2011) NCBI GEO: archive for functional genomics data sets-10 years on. Nucleic Acids Res 39:1005-1010.
2. Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, et al. (2013) Array Express update-trends in database growth and links to data analysis tools. Nucleic Acids Res 41: 987-990.
3. Larsson O, Sandberg R (2006) Lack of correct data format and comparability limits future integrative microarray research. Nat Biotechnol 24: 1322-1323.
4. Larsson O, Wennmalm K, Sandberg R (2006) Comparative microarray analysis. Omics 10: 381-397.
5. Moreau Y, Aerts S, De Moor B, De Strooper B, Dabrowski M (2003) Comparison and meta-analysis of microarray data: from the bench to the computer desk. Trends Genet 19: 570-577.
6. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res 62: 4427-4433.
7. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, et al. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 29: 365-371.
8. Witwer KW (2013) Data submission and quality in microarray-based microRNA profiling. Clin Chem 59: 392-400.
9. Xia J, Gill EE, Hancock RE (2015) NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. Nat Protoc 10: 823-844.
10. Novianti PW, Tweel IV, Jong VL, Roes KCB, Eijkemans MJC (2015) An Application of Sequential Meta-Analysis to Gene Expression Studies. Cancer Inform 14: 1-10.
11. Nguyen T, Tagett R, Donato M, Mitrea C, Draghici S (2015) A novel bi-level meta-analysis approach: applied to biological pathway analysis. Bioinformatics.
12. Hughey JJ, Butte AJ (2015) Robust meta-analysis of gene expression using the elastic net. Nucleic Acids Res 43: 79.