ORIGINAL ARTICLE

A framework for systematic reviews and meta-analyses of high throughput gene expression datasets: A review on meta-analysis tools with no coding skills

ELHAM AMJAD[1**], SOLMAZ ASNAASHARI[1**], BABAK SOKOUTI[1*]

[1] *Biotechnology Research Center, Tabriz University of Medical Sciences, Tabriz, Iran*

**Correspondence to Babak Sokouti, Biotechnology Research Center, Tabriz University of Medical Sciences, Tabriz, Iran; Email: b.sokouti@gmail.com; sokoutib@tbzmed.ac.ir; Tel: +98 (41) 3336 40 38, Fax: +98 (41) 3336 40 38.*

ABSTRACT

The significant role of review studies carried out based on systematic review and meta-analysis especially for randomized clinical trials has been frequently met in the literature. In the current review, the required steps and guides for performing a systematic review and meta-analysis on publicly available gene expression omnibus (GEO) repository database are pointed out. Some of the online available tools for performing this type of meta-analysis are introduced and their various features have been demonstrated and discussed. Finally, it has been concluded that the knowledge of robust biomarker discovery in different types of diseases among several species can benefit from systematically reviewing and screening the curated gene expression datasets and meta-analysis approach.

Keywords: Systematic review; Meta-analysis; GEO dataset; Online web tools; Gene expression.

INTRODUCTION

The current study reviewed the required steps and guides for performing a systematic review and meta-analysis on publicly available gene expression omnibus (GEO) repository database. Recently in 2016, a comprehensive and thorough editorial was published on demonstrating the primary and critical requirements for performing systematic reviews and meta-analyses based on gene association based studies[1]. The approach aims to identify potential and significant gene biomarkers between two groups using a set of statistical methods covered in the meta-analysis procedure based on retrospective published gene set enrichment (GSE) datasets. Additionally, two popular and recent interactive online tools, and one automated analytical workflow will be recommendable.

Systematic reviews and meta-analyses: The rationale behind the increasing number of studies carrying out based on systematic reviews and meta-analyses in the past decades is in direct relationship with the exponentially growing trend of the genomic data generated through the use of high-throughput techniques such as microarrays or mass spectrometry, to mention a few for genomics and epigenomics studies2[,3]. For saving, sharing, reusing, and manipulating such genomic data, in the Big Data era, the most well-known public repository database, NCBI's Gene Expression Omnibus (GEO)[4], is one of the first options for biomedical-based and biomarker discovery investigations[2,3]. Currently, the NCBI GEO dataset comprises of 121,659 series, 20,434 platforms, and 3,328,150 samples accounted for 4,745 organisms of which RNA or ArrayExpress (1,565,503) and Sequence Read Archive (SRA: 1,299,594) contributed to the highest counts among other samples types (https://www.ncbi.nlm.nih.gov/geo/summary/).

Available checklists: Having in mind that there are different types of reviews such as narrative, critical, literature, rapid, systematic, and state-of-art [5], among them, the systematic review can be considered as a quantitative and bias-free type once a user has conducted it correctly based on the documented statements (Preferred Reporting Items for Systematic reviews and Meta-analyses (PRISMA) [6] or Meta-analysis Of Observational Studies in Epidemiology (MOOSE)[7] and quality control process of studies [8-10]. However, the unpublished studies with or without negative results may still play a vital role in exerting possible biases (specifically the publication bias) in the outcome of the quantified systematic review[11].

Importance of gene expression of datasets: Taking into account that retrospective and independent studies may only have standalone impacts on the scientific community considering their outcomes, the meta-analysis approach as a set of statistical tools, can be employed to accumulate those retrospective data and derive a quantitative and descriptive result. Since in a meta-analysis procedure, the safety and efficacy of clinical treatments and the trueness and correctness of reported data for a disease are significant factors, so in a clinical research, the studies based on randomized controlled trials (RCTs), as a standard approach used by Cochrane Organization, are highly recommended to obtain the most reliable outcomes[12,13,14]. Surprisingly, several studies based on systematic reviews and meta-analyses have recently extended their subject into the genome era covering both the genome-wide association studies (GWAS) and gene expression omnibus (GEO) datasets[15,16,17]. However, a well-known repository for GEO datasets is the NCBI in which the authors register and upload their datasets using the default form designed by NCBI, which may be lack coherence, especially in terms of knowledge-based descriptions and unified types of samples to mention a few.

A systematic review and meta-analysis framework for GEO datasets: Previously, Lovell proposed comprehensive guidance for systematic reviews and meta-analyses on GWAS[1]. Besides the GWAS, it is essential to account for demonstrating an outline for GEO datasets of the specific disease. This outline includes *(i)* dataset resource, *(ii)* purpose or question of research, *(iii)* searching date, *(iv)* PRISMA flowchart for systematic review procedure, *(v)* quality check, *(vi)* minimum requirements, and *(vii)* available meta-analysis tools.

Database source – One of the most well-known repository databases that host the required genome datasets is the NCBI GEO database (i.e., https://www.ncbi.nlm.nih.gov/gds).

Purpose – The question of the research is defined based on the specific diseases or the effects of pharmaceutical agents in which several gene biomarkers (among ~33000 probeset IDs within one sample) or signaling pathways may play roles.

Searching date – For future scientists and researchers, the searching date is an essential factor to mention; so, if an additional dataset was available, it could be used for an updated meta-analysis assessment.

PRISMA flow diagram – The diagram flow starts from searching a genome dataset resource, continuing with the inclusion and exclusion criteria as well as reaching the final remaining dataset for further analysis. The arrangement of inclusion and exclusion criteria in the systematic review process can preserve the homogeneity of the datasets by considering the same contents in terms of source of extraction (e.g., lung tissue, peripheral blood mononuclear cell (PBMC)) and extraction protocol (e.g., mRNA, miRNA) as well as the sample types (e.g., healthy *vs.* lung cancer). However, different platforms (e.g., GPL570, GPL96) and file types (CEL, TXT) can be involved. Moreover, some of the meta-analysis tools work with only CEL files, and others mostly depend on the series matrix file(s). Specifically, one may also take advantage of the meta-analysis tools to investigate the involved gene biomarkers between two or more different diseases and species.

Quality check – The quality of the datasets should be reviewed and scored to prevent the potential biases regarding the samples (i.e., the intensities assigned for probeset IDs of each class type) within the included datasets. The useful recommendation is that the quality of samples needed to be assessed based on a modified standard deviation *vs.* the median value where the SD values greater than 0.3 would be of low quality[18].

The minimum requirements for the number of datasets is two, which were screened and reviewed by at least two independent researchers. Meta-analysis tools- To carry out the final step, the meta-analysis process, there are two popular online tools, namely ExAtlas[18,19] and ImaGEO (Integrative Gene Expression Meta-Analysis from GEO database)[20], as well as a standalone MAAMD (Meta-Analysis of Affymetrix Microarray Data analysis) schematic programming workflow[21,22]. In the following, the descriptions of the unique characteristics and properties related to each of the meta-analysis tools developed explicitly for identifying candidate biomarkers from the gene expression datasets are in more detail.

Meta-analysis tools in-depth: The MAAMD workflow designed in the Kepler environment has incorporated R free software environment with R-based Bioconductor packages and Python-based AltAnalyze to perform the meta-analysis by running intra-/inter-dataset comparisons without any special computer skills[21,22]. The standardized workflow runs the meta-analysis procedure using some .csv files as input (include the information required for samples, datasets (usually Affymetrix microarray CEL files) and local locations), performs the quality control of datasets, DEG analysis, and then, makes the comparisons between the datasets in terms of experiments or species[21,22]. The ImaGEO online web tool, implemented in open source deployed Shiny server, which is the Rstudio platform to host a Shiny app, was developed by Toro-Domínguez et al[20].. The "ImaGEO" is capable of running the meta-analysis process after checking the data quality between two sets of experiments by using or uploading the microarray datasets from which the samples can be selected and assigned to the specific groups, including control and case. Finally, the web tool generates the required R codes (including all the functions and methods to carry out the analysis) for locally future exporting the results without the need for using the web tool again. The ExAtlas, an online web tool[18,19], is also downloadable to make a CGI based web server work locally within the Microsoft Windows products (using the XAMPP server available at https://www.apachefriends.org/index.html) or Linux (e.g., CentOS, Debian, Ubuntu). The dataset entry of ExAtlas comprises three ways (embedded public database, search in the GEO database, and upload the gene expression profile matrix text files). The ExAtlas takes advantage of four meta-analysis methods (Fisher's, z-score, fixed effect, and random effect) between pair of experiment groups (i.e., case and control) in which usage of same methodologies for datasets among different or same species are useful[19]. However, the results retrieved for the random-effects model can be beneficial when different methods are also involved, and in using different species, it will only compare their orthologs. Moreover, the quality control of the datasets can be carried out before meta-analysis procedure to remove those samples or datasets with lower quality SD values (SD > 0.3 is a criterion for being determined as low quality). Additionally, the ExAtlas has user credential based profiles to save the obtained results for future extraction and alterations.

DISCUSSION

In contrast to the standard meta-analysis of genome-wide association studies, the current meta-analysis is not prone to several well-known biases such as publication bias, selection bias, selective bias, diagnosis bias. Due to the fact the registration of microarray datasets is compulsory before the corresponding original article published, the effects of negative results, whether they are published or remained unpublished, will not exist. Moreover, in the current types of meta-analysis investigations, target gene biomarkers will be well determined among the included studies and analyzed in terms of their functional enrichment pathways. But in the standard types, the target genes are known before performing the systematic review and meta-analysis. Typically without including the systematic review and meta-analysis studies for robust biomarker identification among gene datasets, the systems biology and bioinformatics approaches are utilized on one or more NCBI-GEO identical datasets in terms of including control and case groups of the same disease[23]. In this approach, the important differentially expressed genes will be determined using statistical methods such as ANOVA by significant p-values and fold changes of about greater than two, and then, it reveals their potential involvement in signaling pathways as well as gene biomarker identification through constructing protein-protein interaction networks and classifying the gene modules.

Considering the ExAtlas online meta-analysis tool, some worked examples are available in the literature to perform a systematic review and meta-analysis on the NCBI-GEO datasets. These researches have worked on NCBI-GEO datasets to identify differentially expressed genes (DEGs) associated with signaling pathways (e.g., estrogen receptor and Wnt) involved in diseases such as cerebral aneurysm and chronic obstructive pulmonary disease[15,24,25] by systematically searching the NCBI-GEO dataset repository and using the remained datasets for meta-analysis approach. The results showed that the identified DEGs were in agreement with experimental and clinical outcomes, however in some cases (e.g., a comparative hypoxia study) carried out across various types of species by using MAAMD workflow, the meta-analysis of these datasets could also determine novel conserved genes[21,22]. Also, a comparative case study between Alzheimer's disease and lung cancer using ImaGEO has revealed deregulated genes between two conditions that were entirely in agreement with the literature outcome[20,26].

CONCLUSION

In conclusion, the current guidelines is the recommended procedure to carry out a systematic review and meta-analysis on high throughput gene expression datasets to identify novel biomarkers in specific disease(s) between/among diverse species.

REFERENCES

1. Lovell DP (2016) systematic reviews and meta-analyses of gene association studies. Biomarkers:673-677
2. Chen G, Ramírez JC, Deng N, Qiu X, Wu C, Zheng WJ, Wu H (2019) Restructured GEO: restructuring Gene Expression Omnibus metadata for genome dynamics analysis. Database 2019. doi:10.1093/database/bay145
3. Wang Z, Lachmann A, Ma'ayan A (2019) Mining data and metadata from the gene expression omnibus. Biophysical reviews 11 (1):103-110
4. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M (2012) NCBI GEO: archive for functional genomics data sets—update. Nucleic acids research 41 (D1):D991-D995
5. Grant MJ, Booth A (2009) A typology of reviews: an analysis of 14 review types and associated methodologies. Health Information & Libraries Journal 26 (2):91-108
6. Moher D, Liberati A, Tetzlaff J, Altman DG (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Annals of internal medicine 151 (4):264-269
7. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, Moher D, Becker BJ, Sipe TA, Thacker SB (2000) Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. Jama 283 (15):2008-2012. doi:10.1001/jama.283.15.2008
8. Bown M, Sutton A (2010) Quality control in systematic reviews and meta-analyses. European Journal of Vascular and Endovascular Surgery 40 (5):669-677
9. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J (2003) The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC medical research methodology 3 (1):25
10. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of internal medicine 155 (8):529-536
11. Mlinarić A, Horvat M, Šupak Smolčić V (2017) Dealing with the positive publication bias: Why you should really publish your negative results. Biochem Med (Zagreb) 27 (3):030201-030201. doi:10.11613/BM.2017.030201
12. Kabisch M, Ruckes C, Seibert-Grafe M, Blettner M (2011) Randomized controlled trials: part 17 of a series on evaluation of scientific publications. Dtsch Arztebl Int 108 (39):663-668. doi:10.3238/arztebl.2011.0663
13. Moher D, Olkin I (1995) Meta-analysis of Randomized Controlled Trials: A Concern for Standards. Jama 274 (24):1962-1964. doi:10.1001/jama.1995.03530240072044
14. da Costa BR, Jüni P (2014) Systematic reviews and meta-analyses of randomized trials: principles and pitfalls. European Heart Journal 35 (47):3336-3345. doi:10.1093/eurheartj/ehu424
15. Lai PMR, Du R (2019) Differentially Expressed Genes Associated with the Estrogen Receptor Pathway in Cerebral Aneurysms. World neurosurgery 126:e557-e563
16. Jiang Q, Sun Y, Liu X (2019) CXCR4 as a prognostic biomarker in gastrointestinal cancer: a meta-analysis. Biomarkers 24 (6):510-516
17. Yap NY, Yap FN, Perumal K, Rajandram R (2019) Circulating adiponectin as a biomarker in renal cell carcinoma: A systematic review and meta-analysis. Biomarkers (just-accepted):1-24
18. Sharov AA, Schlessinger D, Ko MS (2015) ExAtlas: An interactive online tool for meta-analysis of gene expression data. Journal of bioinformatics and computational biology 13 (06):1550019
19. Sharov AA, Schlessinger D (2018) ExAtlas: Online Tool to Integrate Gene Expression and Gene Set Enrichment Analyses. In: Molecular-Genetic and Statistical Techniques for Behavioral and Neural Research. Elsevier, pp 73-104

20. Toro-Domínguez D, Martorell-Marugán J, López-Domínguez R, García-Moreno A, González-Rumayor V, Alarcón-Riquelme ME, Carmona-Sáez P (2018) ImaGEO: integrative gene expression meta-analysis from GEO database. Bioinformatics 35 (5):880-882
21. Gan Z, Wang J, Salomonis N, Altintas I, McCulloch AD, Zambon A MAAMD: A Workflow to Standardize Meta-Analyses of Affymetrix Microarray Data. In: 2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology, 2012. IEEE, pp 120-120
22. Gan Z, Wang J, Salomonis N, Stowe JC, Haddad GG, McCulloch AD, Altintas I, Zambon AC (2014) MAAMD: a workflow to standardize meta-analyses and comparison of affymetrix microarray data. BMC bioinformatics 15 (1):69
23. Amjad E, Asnaashari S, Sokouti B, Dastmalchi S (2020) Systems biology comprehensive analysis on breast cancer for identification of key gene modules and genes associated with TNM-based clinical stages. Scientific reports 10:39315. https://doi.org/10.1038/s41598-020-67643-w
24. Amjad E, Asnaashari S, Sokouti B (2020) The role of associated genes of Wnt signaling pathway in chronic obstructive pulmonary disease (COPD). Gene Reports 18:100582. doi:https://doi.org/10.1016/j.genrep.2019.100582
25. Amjad E, Asnaashari S, Sokouti B (2020) Induction of mTOR signaling pathway in the progression of papillary thyroid cancer. Meta Gene 24:100704. doi:https://doi.org/10.1016/j.mgene.2020.100704
26. Sánchez-Valle J, Tejero H, Ibáñez K, Portero JL, Krallinger M, Al-Shahrour F, Tabarés-Seisdedos R, Baudot A, Valencia A (2017) A molecular hypothesis to explain direct and inverse co-morbidities between Alzheimer’s Disease, Glioblastoma and Lung cancer. Scientific reports 7 (1):4474.