Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Feature-expression heat maps – A new visual method to explore complex associations between two variable sets

Bartholomeus C.M. (Benno) Haarman [a,*], Rixt F. Riemersma-Van der Lek [a], Willem A. Nolen [a], R. Mendes [b], Hemmo A. Drexhage [c], Huibert Burger [a,d]

[a] University of Groningen, University Medical Center Groningen, Department of Psychiatry, Groningen, The Netherlands
[b] Health E-Solutions, Rotterdam, The Netherlands
[c] Erasmus MC, Rotterdam, Department of Immunology, The Netherlands
[d] University of Groningen, University Medical Center Groningen, Department of General Practice, Groningen, The Netherlands

ARTICLE INFO

Article history:
Received 1 May 2014
Accepted 7 October 2014
Available online xxxx

Keywords:
Graph
Heat map
Method
Phenotype
Genotype
Associations

ABSTRACT

Introduction: Existing methods such as correlation plots and cluster heat maps are insufficient in the visual exploration of multiple associations between genetics and phenotype, which is of importance to achieve a better understanding of the pathophysiology of psychiatric and other illnesses. The implementation of a combined presentation of effect size and statistical significance in a graphical method, added to the ordering of the variables based on the effect-ordered data display principle was deemed useful by the authors to facilitate in the process of recognizing meaningful patterns in these associations.
Materials and methods: The requirements, analyses and graphical presentation of the feature-expression heat map are described. The graphs display associations of two sets of ordered variables where a one-way direction is assumed. The associations are depicted as circles representing a combination of effect size (color) and statistical significance (radius).
Results: An example dataset is presented and relation to other methods, limitations, areas of application and possible future enhancements are discussed.
Conclusion: The feature-expression heat map is a useful graphical instrument to explore associations in complex biological systems where one-way direction is assumed, such as genotype-phenotype pathophysiological models.

1. Introduction

Tukey emphasized that exploratory data analysis relies more on graphical display, whereas confirmatory data analysis is easier to computerize [1,2]. Existing graphical methods to explore associations in a set of multiple variables are cluster heat maps and correlation plots. Heat maps originated from two-dimensional displays of a two-by-two data matrix. Larger values were represented by darker squares and smaller values by lighter squares [3]. E.g., in gene expression studies, these values correspond to the amount of a particular RNA or protein expressed. The further development of the cluster heat map, which includes ordering of the columns and rows to reveal structure, has been a multi-step process. Facilitating the process of detecting meaningful patterns in the visual presentation, Sneath [4] displayed the results of a cluster analysis by permuting the rows and the columns of a matrix to place similar values adjacent to each other according to the clustering, which is based on the *effect-ordered data display* principle [5]. This principle says that in any data table or graph, unordered variables should be ordered according to what we aim to show. The ideas of similarity and grouping are derived from Gestalt psychology, but have shown to be equally useful in biology [6]. Ling ultimately formed the idea for joining cluster trees to the rows and columns of the heat map [7]. Technical advances in printing let the presentation of the graphs develop from overstruck printer characters to the use of computer programs to produce cluster heat maps with high-resolution color graphics [8], as can be seen in Fig. 1.

Abbreviations: FDR, False Discovery Rate.

* Corresponding author at: Department of Psychiatry, CC44, University Medical Center Groningen, Hanzeplein 1, Postbus 30.001, 9700 RB Groningen, The Netherlands.

E-mail addresses: b.c.m.haarman@umcg.nl (B.C.M. (Benno) Haarman), r.f.riemersma@umcg.nl (R.F. Riemersma-Van der Lek), w.a.nolen@umcg.nl (W.A. Nolen), r.mendes@healthesolutions.nl (R. Mendes), h.drexhage@erasmusmc.nl (H.A. Drexhage), h.burger@umcg.nl (H. Burger).

http://dx.doi.org/10.1016/j.jbi.2014.10.003
1532-0464/

Correlation plots are used to visualize association matrices. These plots can be regarded as heat map style displays of multiple correlation statistics. These statistics may be drawn in several forms: as numbers, circles, ellipses, squares, bars or "pac-man" symbols. In each symbol both the sign and magnitude of the correlation coefficient is represented. This is done so by using two colors printed with varying intensity. The color indicates the sign of the coefficient and the intensity of the color increases proportionally with the magnitude of the correlation coefficient [5].

We entertained the idea that these visual methods could be of help in the exploration of associations between genetic data and phenotypical presentation in the investigation of the pathophysiology of psychiatric disorders. The pathophysiology of these disorders is still largely unknown. In an effort to unravel the genetic basis of mood disorders, many genome-wide association studies have been performed. However, these studies found evidence for only a few susceptibility genes, which in turn accounted for a very minor part of disease liability. This fuelled the idea that to grasp the mechanism of these complex illness, it is important to have a framework integrating biology and clinical phenotype [9]. In this model the intermediary processes that occur between the genetic information and the specific phenotypical expression of these illnesses are regarded as a *black box* [10].

To achieve a better understanding of these intermediary underlying pathophysiological processes, we wanted to investigate patterns in the associations between specific symptoms and specific gene expression [11]. We hypothesized that patterns in these associations would come to light most effectively at the intersection of related genes and related symptoms, embroidering on the above mentioned principle of *effect-ordered data display*.

Because we were exploring the physiology of these intermediary *black box* processes, we preferred to use an effect size measure instead of the correlation coefficient. Contrary to the correlation coefficient effect size measures describe the magnitude of an association in measurement units. This is generally of more interest in the biomedical sciences than just the degree of linearity of an association, which is measured by the correlation coefficient. This is of special importance in explorative biology based research, which can be compared to a field biologist visiting a new habitat who will begin describing the most striking features, i.e. analogous to the largest effects sizes. In addition to a measure of the magnitude of the associations of interest we wanted to implement inferential statistics to aid in drawing conclusions incorporating their certainty. Statistical significance for the given sample size was used in this regard.

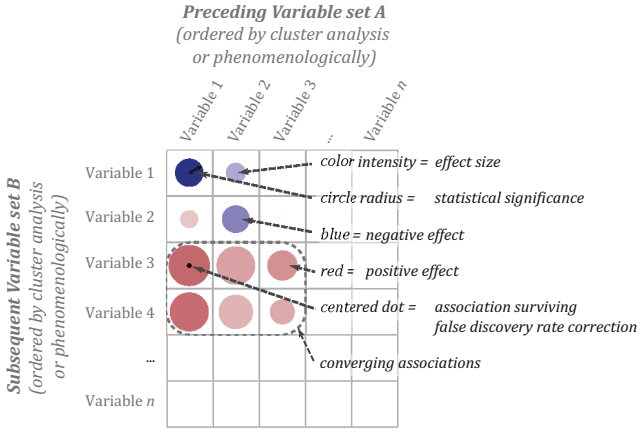

Fig. 2. Overview of a feature-expression heat map.

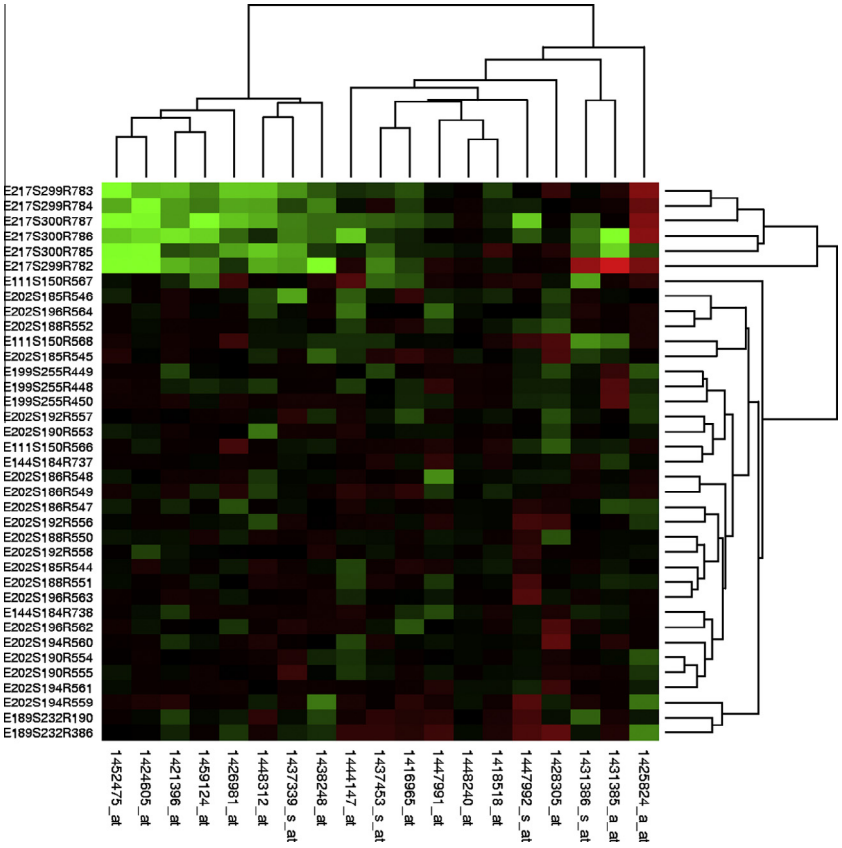

Fig. 1. Cluster heat map [27,28]. The columns of the heat map represent genes and the rows represent samples. Each cell is colorized based on the level of expression of that gene in that sample.

Please cite this article in press as: Haarman BCM(et al. Feature-expression heat maps – A new visual method to explore complex associations between two variable sets. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.10.003