Journal of Biomedical Informatics 53 (2015) 156–161

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 14 October 2014

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

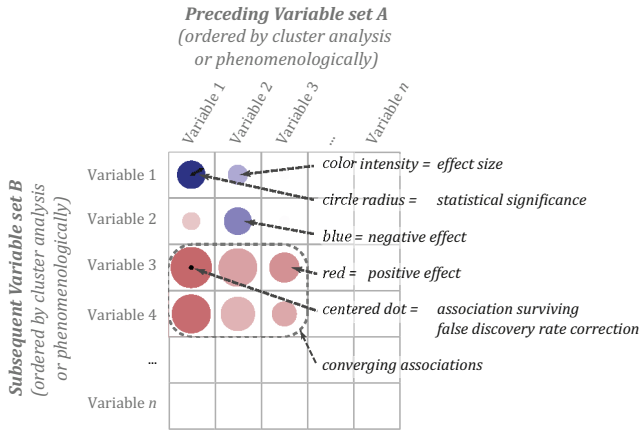

Fig. 2. Overview of a feature-expression heat map.

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.

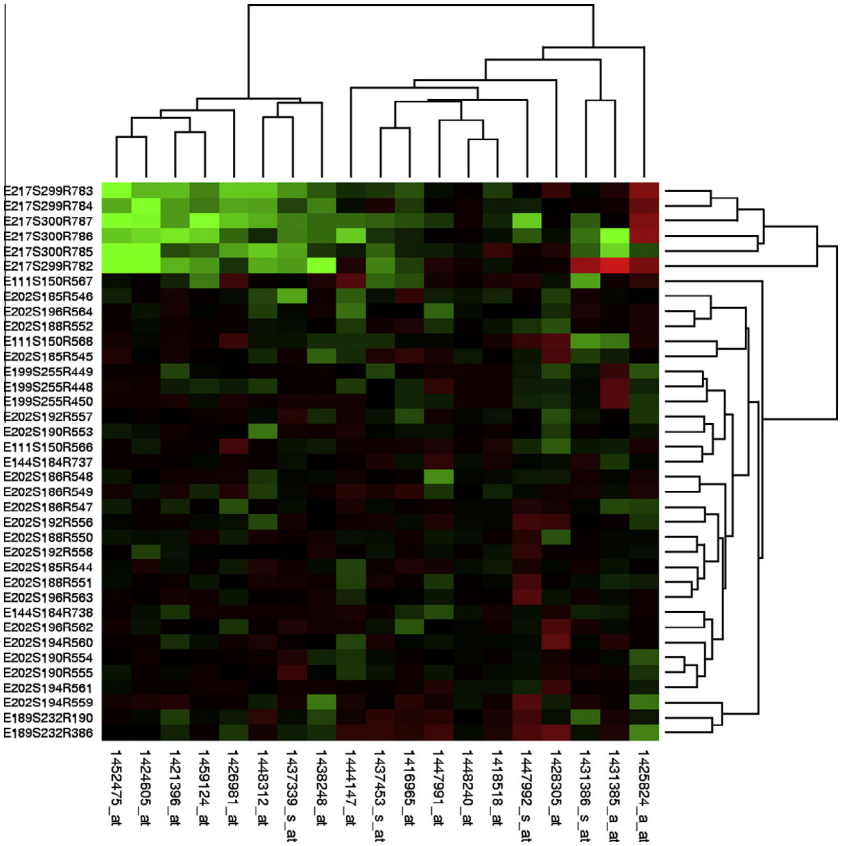

Fig. 1. Cluster heat map [27,28]. The columns of the heat map represent genes and the rows represent samples. Each cell is colorized based on the level of expression of that gene in that sample.

Summarizing, the scope of the method we had in mind was to visualize a large set of associations of variables in two sets in which one-way association was assumed, i.e. from gene expression to phenotype. This approach required a contiguously and ordered arrangement of the variables, incorporating the direction of the associations, an effect size measure and the statistical significance of individual associations. Elaborating on this reasoning we developed feature-expression heat maps.

In this article we will first describe the method of creating a feature-expression heat map. Secondly we will present an example. Finally we will comment on this method and theorize on other areas of application.

2. Materials and methods

2.1. Preprocessing

The dataset for a feature-expression heat map analysis needs to consist of two sets of variables that differ in their nature. The variables in these sets are assumed to represent phenomena that occur in a certain time order according to an underlying theoretical model and have a one-way relationship. The variables of each set are then transformed in a way that facilitates comparability within the set, such as *Z*-transformation. Because of reasons of in-between comparability, similarity in data type (binary, ordinal, continuous) of the variables within each set is a stringent requirement.

In order to be able to recognize meaningful patterns in the final feature-expression heat map the variables need to be arranged contiguously and ordered *a priori* in a way that similar variables are placed adjacent to each other in both variable sets, consistent with the *effect ordered data display* principle. This may be achieved by performing a cluster analysis on a correlation matrix of the variables in each variable set. Alternatively, the variables can be arranged on phenomenological similarity.

2.2. Analysis

Regression methods may be used to determine the effect size and statistical significance of the associations of the individual variables of the preceding with those of the subsequent variable sets. In case of linear regression the regression coefficient β and its *p*-value are used as measures of effect size and statistical significance, respectively. In case of (ordered) logistic regression the β is used which indicates the change in the logit (or log-odds of the outcome) and may be preferred over the odds ratio, the latter being asymmetric and ranging from zero to infinity.

In order to control for the increased risk of wrongful rejection of the null hypothesis (type I error, false positive results) correction for the false discovery rate (FDR) can be applied, as described by Benjamini and Hochberg [12]. Deriving from this method approximations of the power and sample size can be calculated [13,14].

The separate effect size, statistical significance and FDR results of each association are put into individual data matrices. They are ordered in such a way that for each association the statistical property is placed on the intersection between the preceding variables (columns) and the subsequent variables (rows), where the variables of the preceding and subsequent variable sets are ordered according to the above-described procedure.

2.3. Graphical presentation

To display the associations of two sets of variables adapted heat maps are drawn, visualizing the preceding variable set in the columns and the subsequent variable set in the rows, each ordered facilitating the visual identification of meaningful clusters of association later on. An underlying cluster analysis tree may be added to one or both of the axes.

In the feature-expression heat maps the associations between preceding and subsequent variables are represented by circles (Fig. 2). The effect size measure is represented by the type and intensity of the color, whereas the statistical significance of the analyses is represented by the radius of the circles. Shades of red are used for positive effect sizes, whereas shades of blue are used for negative effect sizes. Centered dots may be added to the compartments that comply with a FDR below a certain threshold, thus allowing for a selected portion to be expected false positive. Optionally, when needed in the process of drawing statistical decisions it can be preferred to only visualize the circles of associations of which the statistical significance is below a pre-defined threshold.

To create these heat maps, separate plots are drawn for the effect size, statistical significance and FDR data matrices. These plots can be drawn with the corrplot package, by Wei [15], on R (R Development Core Team 2013, Vienna, Austria) [16]. R programming code examples are given in Table 1. In order to magnify the more significant associations, i.e. those in which p is approaching 0, applying a transformation $1 - \sqrt[3]{p}$ to the statistical significance parameter was found to be useful empirically. Finally, the separate plots can be merged with the transparency function of a vector graphics editor.

Table 1
R code for corrplot.

Step	Code

Import tables as matrices	`> plot_size <- as.matrix(read.table(''plot_size.txt'', sep = ''\t'', header = TRUE))`
	`> plot_significance <- as.matrix(read.table(''plot_significance.txt'', sep = ''\t'', header = TRUE))`
	`> fdr_template <- as.matrix(read.table(''fdr_template.txt'', sep = ''\t'', header = TRUE))`
	`> fdr_significance <- as.matrix(read.table(''fdr_significance.txt'', sep = ''\t'', header = TRUE))`

Define colors	`> col <- colorRampPalette(c(''blue'', ''white'', ''red''))`
Create significance plot	`> corrplot(plot_significance, method =c(''circle''), col=(''black''), tl.cex = 0.6, tl.col = (''black''), cl.pos=''n'')`
Create size measure plot	`> corrplot(plot_size, is.corr = FALSE, method =c(''color''), addgrid.col = ''grey'', col = col(200), tl.cex = 0.6, tl.col = ''black'', cl.pos=''n'')`
Create FDR plot	`> corrplot(fdr_template, method =c(''circle''), col=(''black''), tl.cex = 0.6, tl.col = (''black''), p.mat = fdr_significance, insig = ''blank'', sig.level = 0.009, cl.pos=''n'')`
Create legends	`> corrplot(legend_significance, method =c(''circle''), col=(''grey''), tl.cex = 0.6, tl.col = (''black''), cl.pos=''n'')` `> corrplot(plot_size, is.corr = FALSE, method =c(''color''), addgrid.col = ''grey'', col = col(200), tl.cex = 0.6, tl.col = ''black'', cl.pos=''r'', cl.lim = c(-3,3), cl.ratio=0.4, cl.length=7)`

R code for corrplot version 0.73 [15].

3. Results and discussion

3.1. Example

We will use some results of our recent paper on monocyte gene expression and psychiatric symptoms of patients with bipolar disorder to demonstrate the use of the feature-expression heat map [11]. In this dataset we analyzed the relation between gene expression and manic symptoms containing information from 64 patients.

According to the underlying pathophysiological model manic symptoms are associated with inflammation related monocyte gene expression, albeit indirectly. The manner of how these processes interact is as yet unknown. Resulting from the pathophysiological model gene expression variables were placed in the preceding variable set, whereas manic symptoms were put in the subsequent variable symptom set. The gene variables were ordered based on a hierarchical cluster analysis of a Pearson correlation matrix, previously executed and published [17]. They were subdivided into three subclusters and a rest group based on the molecular function. Symptom variables were ordered into a phenomenological sequence ranging from core mood symptoms via thought symptoms, psychosomatic symptoms, motor symptoms, food intake symptoms, sleep symptoms, to higher functional symptoms. The gene expression variables contained continuous, normally distributed data and *Z*-transformation was applied. The symptom variables contained ordered categorical data, which were all transformed on a 0–1 scale for standardization.

In this example analysis associations between individual gene expression and individual manic symptoms were analyzed using ordered logistic regression. The ordered logistic regression model is a direct generalization of the commonly used two-outcome logistic model. In ordered logistic regression, an ordinal dependent variable is estimated as a linear function of independent variables and a set of cut-points [18,19].

For each association used to create the feature-expression heat map in this example the effect size was defined as the magnitude of the regression coefficient β indicating the change in the log-odds of the outcome variable per unit increase in the independent variable.

Subsequently, the *Z*-statistic was calculated for each association as the ratio of the coefficient β to its standard error. The statistical significance then was defined as the statistical probability (p) that this statistic, assumed to follow a normal distribution, is as extreme as, or more so, than what would have been observed under the null hypothesis, defined by $p > |Z|$.

As stated previously, the transformation $1 - \sqrt[3]{p}$ was applied to the statistical significance and maximization of the effect size to 3 was applied. Effect size and significance plots were exported as vector graphs, displayed in Fig. 3a and b.

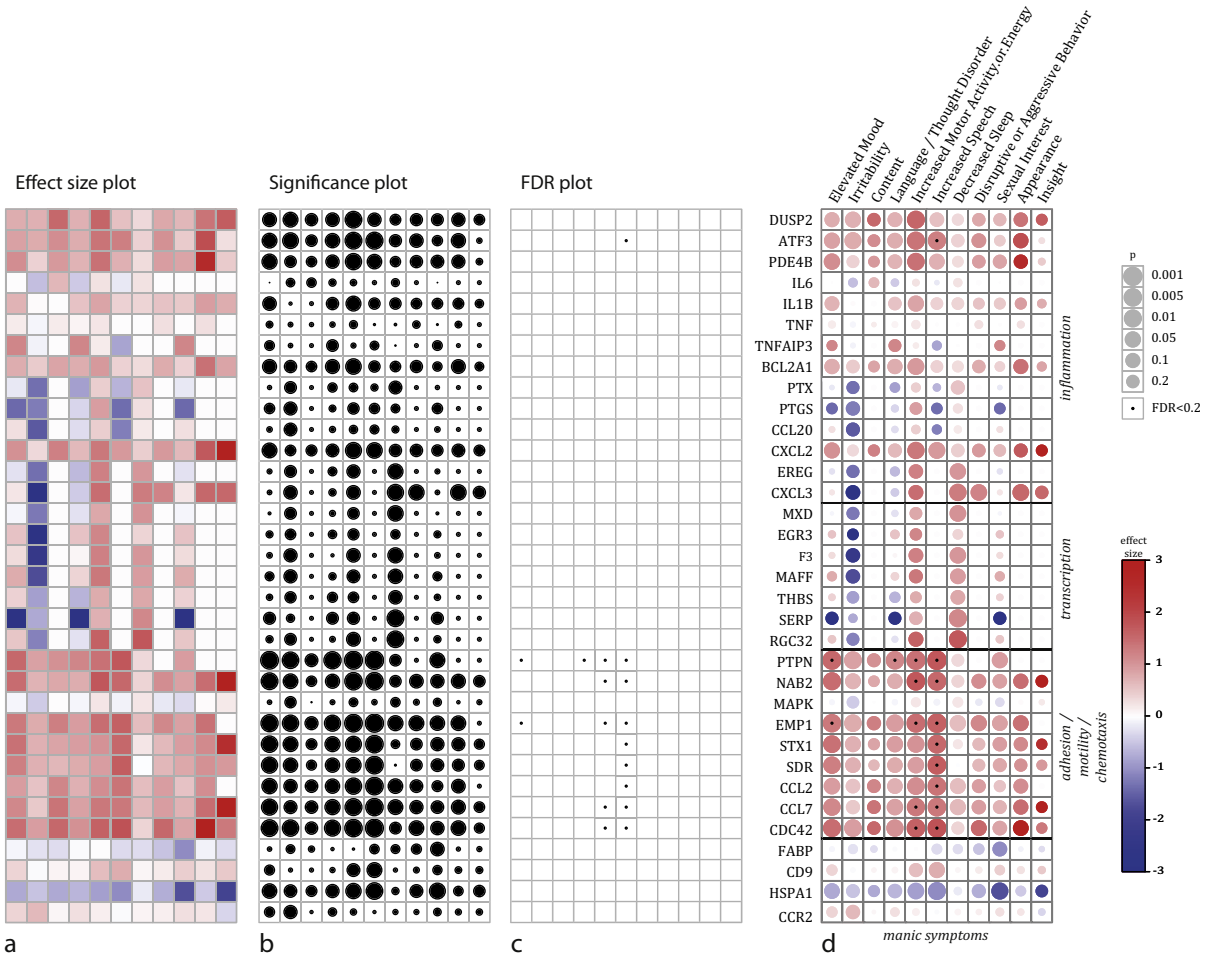

Fig. 3. Individual effect size plot (a), statistical significance plot (b) and FDR plot (c) that constitute the feature-expression heat map (d) depicting the associations between manic symptoms and monocyte inflammatory gene expression [11]. Manic symptoms were ordinally measured. Gene expression was normally distributed and z-transformed. Statistical analysis was performed using ordered logistic regression. Circles with a center dot represent significance below the 0.2 false discovery rate (FDR) threshold for multiple testing. The legends, black lines and annotations have been added manually after the creation of the heat map. Fig. 3d has been reprinted with permission [11].

We set the FDR at 0.2, thus allowing 1/5 to be false positive. This resulted in a statistical significance threshold (q-value) of 0.009. A FDR plot was drawn with corrplot marking the corresponding circles (Fig. 3c). Finally the effect size plot, statistical significance plot and FDR plot were merged with Adobe Illustrator (Adobe Systems Inc., San José, California; Fig. 3d).

This mania feature-expression heat map allowed for the identification of a converging group of associations between the genes PTPN – CDC42, so called sub-cluster 2 genes, and manic symptoms. Sixteen of these associations were significant after FDR correction, especially in the associations with the symptoms increased speech and increased motor activity [11].

3.2. Relation to other methods

Marked properties of the feature-expression heat maps are the combined display of an effect size measure and the statistical significance and use of *effect-ordered data display* on two sets of variables. This combination aids in the recognition of association patterns in complex systems, e.g. pathophysiological models.

These feature-expression heat maps are based on the original cluster heat maps and use some features that can be found in correlation plots. Both the original cluster heat maps and feature-expression heat maps facilitate the visual analysis of extensive data sets for patterns. Where original cluster heat maps allow displaying all kinds of data matrices, the feature-expression heat map limits its applicability to one-way associations between two variable sets. While limiting the area of usability, it facilitates the use of regression methods. These are almost essential to analyze the strength of the phenomena involved and are an asset in explorative research focused on deconstructing pathophysiological models.

Of the various ways of displaying correlation plots the circle correlation plot, where the radius as well as the type and intensity of the color were derived from the correlation coefficient, drew our attention. In doing so the correlation plot utilizes two ways of visual display to show one test outcome. The visual combination of two measures of association in the feature-expression heat maps, i.e. effect size and statistical significance in one graphical display, increases the usefulness of the method. While visualizing the effect size allows observation of the strongest associations, adding the significance of the association adds the ability to observe the signal-to-noise ratio of the observations, thus aiding in the process of inference of the explorative process. In visually integrating and presenting this distinctive information this method allows for a balanced interpretation of the associations. Especially, the method allows for salient complex (patterns of) associations to become apparent.

Bipartite network graphs can be regarded as an alternative to the feature-expression heat map method in visualizing two dimension (bipartite) variables [20,21]. For example, a bipartite network could represent genes and symptoms as nodes, and the edge weight connecting the nodes could represent the significance of the respective association. The color of the edges could represent the effect size, and its style (e.g. solid versus dotted line) could represent whether or not the significance met the FDR correction. The possibility of adding more than two variable dimensions in one graph (multipartite graph) is a benefit of this method. However, in medium size datasets consisting of two variable dimensions the authors consider the surveyability of the feature-expression heat map to be favorable due to its convenient arrangement.

3.3. Limitations

The scalability of the feature-expression heat map is principally limited by the perception and interpretation capabilities of the interpreter. To facilitate the interpretability of more complex feature-expression heat maps separate panels can be created containing subgroups of the variables. For example, the presentation of the separate symptom dimensions (manic, depressive, psychotic symptoms) has been performed using separate panels in our recent study [11].

Like all methods exploring data sets with large variable lists, feature-expression heat maps may bring about an increased risk of type I errors. Although an extensive discussion about controlling for this problem is beyond the scope of this article, we have endeavored to restrict the extend of this limitation by deploying the Benjamini-Hochberg method. It is known that the Benjamini-Hochberg method offers a more powerful alternative to the traditional Bonferroni method [22,23]. Although a wrongful rejection of the null hypotheses cannot be fully eliminated with this method, considering only clusters of adjoining associations to be meaningful can further diminish this risk, which is a strength of the heat map method.

3.4. Future perspectives

Originating in a study on the relation between monocyte gene expression and psychiatric symptoms we expect the feature-expression heat map method to be useful for studying many other illnesses by benefiting from a combined effect size and statistical significance plot. High throughput screening studies [24,25] involving complex relations between genetics and biological features are obvious candidates. Even more so, the method is not limited to genotype–phenotype relations, but can easily be applied in any explorative analysis exploring multiple variables, within a group of subjects, of which a two-variable-set one-way dependency is assumed.

As this method can be regarded as an evolution of the original cluster heat maps, it is tempting to reflect on possible future enhancements. At present the method relies on separate ordering of the row and column variables, based on individual cluster analyses or phenomenological similarity. Biclustering is a cluster method that allows simultaneous clustering of both rows and columns [26]. Biclustering of the data matrix to obtain the ordering of the variables would increase the extent of the interaction effects in the feature-expression heat map. Furthermore, instead of relying on visual identification of meaningful clusters of associations, future development incorporating more advanced, automated pattern recognition may aid in the discrimination between more meaningful and less meaningful clusters. By automation assigned cluster associations could be visually marked by a distinguishable background color of the compartments involved or with a bold line surrounding these compartments.

4. 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. It utilizes the combined display of an effect size measure and the statistical significance as well as the use of *effect-ordered data display* of two sets of variables, both aiding in the recognition of meaningful association patterns.

Acknowledgments

This study was funded in part by EU-FP7-HEALTH-F2-2008-222963 'MOODINFLAME'. The funding organization had no further role in the study design; collection, analysis and interpretation of

data, the writing of the report and the decision to submit the paper for publication.

Appendix A. Supplementary material

Supplementary material associated with this article consisting of an R package with the example data that has been presented and sample code to generate the individual plots can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2014.10.003.

References

[1] Tukey JW. Exploratory data analysis. Addison-Wesley; 1977.
[2] Tukey JW. We need both exploratory and confirmatory. Am Stat 1980;34:23–5.
[3] Loua T. Atlas statistique de la population de Paris. J. Dejey & cie 1873.
[4] Sneath PH. The application of computers to taxonomy. J Gen Microbiol 1957;17:201–26.
[5] Friendly M. Corrgrams: exploratory displays for correlation matrices. Am Stat 2002;56:316–24.
[6] Friendly M, Kwan E. Effect ordering for data displays. Comput Stat Data Anal 2003;43:509–39.
[7] Ling RL. A computer generated aid for cluster analysis. Commun ACM 1973;16:355–61.
[8] Wilkinson L, Friendly M. The history of the cluster heat map. Am Stat 2009; 63:179–84.
[9] Schulze TG. Genetic research into bipolar disorder: the need for a research framework that integrates sophisticated molecular biology and clinically informed phenotype characterization. Psychiatr Clin North Am 2010;33: 67–82.
[10] Cauer W. Theorie der linearen Wechselstromschaltungen, vol. I. Leipzig: Akad. Verlags-Gesellschaft Becker und Erler; 1941.
[11] Haarman BCM, Riemersma-Van der Lek RF, Burger H, Netkova M, Drexhage RC, Bootsman F, et al. Relationship between clinical features and inflammation-related monocyte gene expression in bipolar disorder – towards a better understanding of psychoimmunological interactions. Bipolar Disord 2014;16: 137–50.
[12] Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med 1990;9:811–8.
[13] Ferreira JA, Zwinderman AH. Approximate power and sample size calculations with the Benjamini–Hochberg method. Int J Biostat 2006:2.
[14] Efron B. Size, power and false discovery rates. Ann Stat 2007;35:1351–77.
[15] Wei T. corrplot: Visualization of a correlation matrix. R Packag Version 073; 2013.
[16] R Core Team. R: A language and environment for statistical computing 2014.
[17] Drexhage RC, van der Heul-Nieuwenhuijsen L, Padmos RC, van Beveren N, Cohen D, Versnel MA, et al. Inflammatory gene expression in monocytes of patients with schizophrenia: overlap and difference with bipolar disorder. A study in naturalistically treated patients. Int J Neuropsychopharmacol 2010;13:1369–81.
[18] Brant R. Assessing proportionality in the proportional odds model for ordinal logistic regression. Biometrics 1990;46:1171–8.
[19] StataCorp. Ordered logistic regression. Stata 13 Base Ref Man. College Station, TX: Stata Press; 2013.
[20] Dulmage AL, Mendelsohn NS. Coverings of bipartite graphs. Can J Math 1958;10:517–34.
[21] Asratian AS, Denley TMJ, Häggkvist R. Bipartite graphs and their applications. Cambridge University Press; 1998.
[22] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B ... 1995.
[23] Noble WS. How does multiple testing correction work? Nat Biotechnol 2009;27:1135–7.
[24] Giuliano KA, Haskins JR, Taylor DL. Advances in high content screening for drug discovery. Assay Drug Dev Technol 2003;1:565–77.
[25] Abraham VC, Taylor DL, Haskins JR. High content screening applied to large-scale cell biology. Trends Biotechnol 2004;22:15–22.
[26] Eren K, Deveci M, Küçüktunç O, Çatalyürek ÜV. A comparative analysis of biclustering algorithms for gene expression data. Brief Bioinform 2013;14: 279–92.
[27] Andrade M. Heatmap. Wikipedia; 2006.
[28] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95:14863–8.