

MetATT: a web-based metabolomics tool for analyzing time-series and two-factor datasets

Jianguo Xia^{1,*}, Igor V. Sinelnikov² and David S. Wishart^{1,2,3}¹Department of Biological Sciences, ²Department of Computing Sciences and ³National Research Council, National Institute for Nanotechnology (NINT), University of Alberta, Edmonton, Alberta, Canada

Associate Editor: Martin Bishop

ABSTRACT

Summary: Time-series and multifactor studies have become increasingly common in metabolomic studies. Common tasks for analyzing data from these relatively complex experiments include identification of major variations associated with each experimental factor, comparison of temporal profiles across different biological conditions, as well as detection and validation of the presence of interactions. Here we introduce MetATT, a web-based tool for time-series and two-factor metabolomic data analysis. MetATT offers a number of complementary approaches including 3D interactive principal component analysis, two-way heatmap visualization, two-way ANOVA, ANOVA-simultaneous component analysis and multivariate empirical Bayes time-series analysis. These procedures are presented through an intuitive web interface. At the end of each session, a detailed analysis report is generated to facilitate understanding of the results.

Availability: Freely available at <http://metatt.metabolomics.ca>

Contact: jianguox@ualberta.ca

Received on May 18, 2011; revised on June 21, 2011; accepted on June 23, 2011

1 INTRODUCTION

Metabolomics involves the study of multiple small-molecule compounds in a biological system. Biofluids (i.e. urine, blood) are by far the most commonly used sample type in metabolomic studies. These samples can be collected conveniently and relatively non-invasively, and are often used in longitudinal studies such as monitoring disease progression, tracking nutritional interventions or observing drug toxicity. Biofluids have also been increasingly used in cross-sectional studies for biomarker discovery. One important lesson learned from these studies is that metabolic profiles of many biofluids are very sensitive to multiple factors such as diet, age, gender, etc. (Psihogios *et al.*, 2008; Slupsky *et al.*, 2007). Therefore, one often needs to consider other factors together with the biological condition of interest during data analysis. Specially designed statistical approaches have been developed to deal with the multifactor issues as well as for time-series studies, such as parallel factor analysis (PARAFAC) (Harshman and Lundy, 1994), ANOVA-simultaneous component analysis (ASCA) (Smilde *et al.*, 2005), multivariate empirical Bayes time-series analysis (MEBA) (Tai and Speed, 2006) and multivariate multiway analysis of multisource data (Huopaniemi *et al.*, 2010). However, most of

these methods are only available in the form of command-line scripts written in MATLAB (The MathWorks Inc., Natick, MA, USA) or R (<http://www.r-project.org/>). Consequently, researchers need to have a deep understanding of the programming languages in order to use these methods. To help overcome this barrier, we have created MetATT, an easy-to-use web-based tool designed to perform common tasks involved in time-series and two-factor metabolomic data analysis, including identification of major patterns associated with main experimental factors, comparison of temporal profiles across biological conditions, as well as detection and validation of the presence of interactions. MetATT is available at <http://metatt.metabolomics.ca>. It has also been incorporated to MetaboAnalyst (Xia *et al.*, 2009, <http://www.metaboanalyst.ca>).

2 METHODS

2.1 Data visualization

To facilitate data exploration and pattern discovery, MetATT offers two visualization approaches—3D interactive PCA (iPCA) and two-way heatmaps. The iPCA supports rotation, zooming and highlighting. Clicking on any data point will display the characteristic variables of the corresponding sample. The heatmap supports a variety of options on clustering, coloring and sorting. Users can adjust the display by using a combination of the available options from the drop-down menus.

2.2 Statistical algorithms

MetATT supports the classical two-way ANOVA and the two recent multivariate inventions—ASCA and MEBA. The two-way ANOVA method supports within- and between-subjects analysis for time-series and independent sample measures, respectively. The ASCA approach first decomposes the variation of the entire data into individual variations induced by each experimental factor and their interactions; it then applies PCA to each variation to get a simple and interpretable result. MetATT also provides ASCA model validation (Vis *et al.*, 2007) and variable selection (Nueda *et al.*, 2007). The MEBA method assesses treatment differences by comparing time-course mean profiles while allowing for variability both within and between time points, thereby reducing false positives and false negatives.

2.3 Implementation of web application

MetATT's web interface was implemented using the JavaServer Faces technology. The 3D interactive visualization tools are based on the LiveGraphics3D technology (Glaab *et al.*, 2010). The statistical algorithms were written in R. MetATT is hosted on GlassFish (version 3.1) using a Linux operating system (Ubuntu 10.04 LTS). The web application has been successfully tested on all major web browsers with a Java plug-in installed.

*To whom correspondence should be addressed.

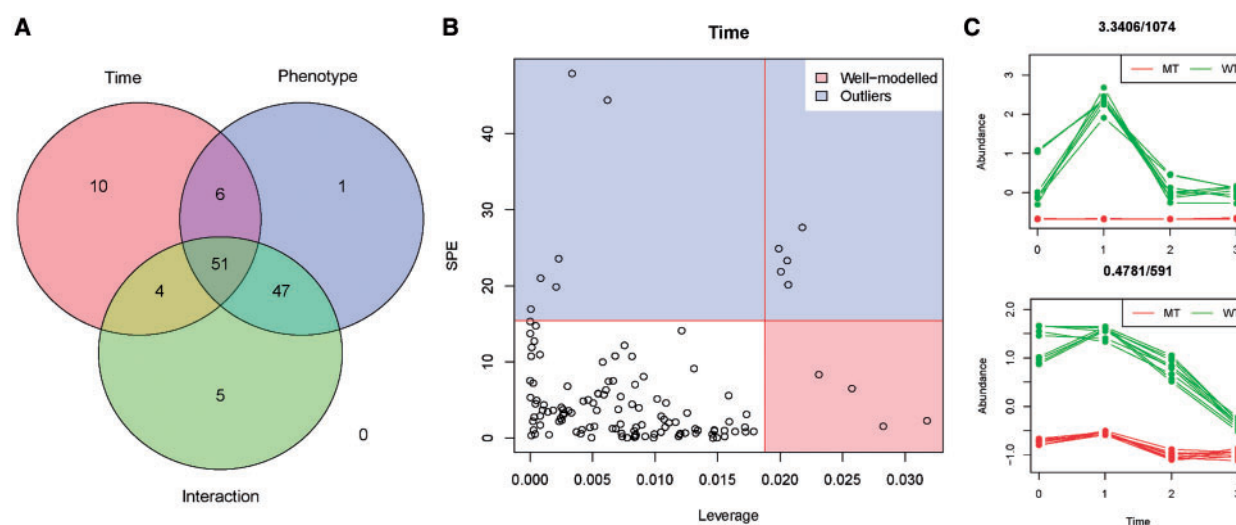


Fig. 1. Some examples of MetATT graphical outputs—(A) Venn diagram summary of results from two-way ANOVA, (B) ASCA selection of important variables associated with Time and (C) two variables with distinctive temporal profiles identified by MEBA.

3 EXAMPLE ANALYSIS

Here, we present the analysis on a subset of the data from a recently published time-course metabolomics study (Meinicke *et al.*, 2008). Accepted data formats are described in the ‘Data Format’ section of the MetATT online documentation. The study was designed to compare the metabolic profiles of two *Arabidopsis thaliana* lines—wide type (WT) and *dde 2-2* mutant (MT) during a wounding time-course. Samples were collected at four time points. This dataset is available as the test data on MetATT’s data upload page. In this example, we use the default parameters for data uploading and normalization.

Data visualization: click the ‘iPCA’ or ‘Heatmap2’ node on the navigation tree on the left panel to access these two methods. Both show clear clustering patterns with regard to time points or phenotypes. **Two-way ANOVA:** click the ‘ANOVA2’ node to access this function. As the data consists of time-series repeated sample measures, the ‘within-subject ANOVA’ is used. The results are summarized in a Venn diagram (Fig. 1A). It is clear that many variables are significant with regard to each experimental effect and their interactions. Click the ‘View details’ link to view a detailed table of these variables. **Detection of major trends:** click the ‘ASCA’ node on the navigation tree. Accept the default parameters for component selection. Click the ‘Major Patterns’ tab to view the major trends associated with each experimental factor as well as the interaction effect. Click the ‘Sig. Features’ tab to view variables that follow these trends (well modeled) as well as those that clearly deviate (outliers) (Fig. 1B).

Testing the presence of interactions: for individual variables, the interaction effect can be assessed by univariate two-way ANOVA. As shown in Figure 1A, the interaction effects are significant for more than half of the variables. The overall interaction effect can be assessed by permutation tests on the ASCA model for interaction. Click the ‘Model Validation’ tab on the ASCA page. The result indicates the overall interaction effect is very significant.

Identification of differential temporal profiles: click the ‘MEBA’ node to perform multivariate empirical Bayes analysis. The result is

a list of all variables ranked by Hotelling’s T^2 . Clicking any variable name will display the corresponding temporal profiles (Fig. 1C).

4 CONCLUSIONS

The growing number of metabolomics being applied to various time-course and multifactor studies has increased the need for user-friendly bioinformatics tools to handle these types of data. MetATT is a full-featured, easy-to-use web-based tool that is designed to perform common tasks for time-series as well as general two-factor metabolomic data analysis and visualization.

Funding: Alberta Innovates Technology Futures; Genome Alberta.

Conflict of Interest: none declared.

REFERENCES

- Glaab,E. *et al.* (2010) vrmigen: an R Package for 3D Data Visualization on the Web. *J. Stat. Softwar.*, **36**, 2347–2348.
- Harshman,R.A. and Lundy,M.E. (1994) PARAFAC: parallel factor analysis. *Comput. Stat. Data Anal.*, **18**, 39–72.
- Huopaniemi,I. *et al.* (2010) Multivariate multi-way analysis of multi-source data. *Bioinformatics*, **26**, i391–i398.
- Meinicke,P. *et al.* (2008) Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps. *Algorithms Mol. Biol.*, **3**, 9.
- Nueda,M.J. *et al.* (2007) Discovering gene expression patterns in time course microarray experiments by ANOVA-SCA. *Bioinformatics*, **23**, 1792–1800.
- Psihogios,N.G. *et al.* (2008) Gender-related and age-related urinalysis of healthy subjects by NMR-based metabolomics. *NMR Biomed.*, **21**, 195–207.
- Slupsky,C.M. *et al.* (2007) Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles. *Anal. Chem.*, **79**, 6995–7004.
- Smilde,A.K. *et al.* (2005) ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics*, **21**, 3043–3048.
- Tai,Y.C. and Speed,T.P. (2006) A multivariate empirical Bayes statistic for replicated microarray time course data. *Ann. Stat.*, **34**, 2387–2412.
- Vis,D.J. *et al.* (2007) Statistical validation of megavariate effects in ASCA. *BMC Bioinformatics*, **8**, 322.
- Xia,J. *et al.* (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res.*, **37**, W652–W660.