Metabolomics 2011   Cairns | Australia

Two-factor and time-series metabolomic data analysis

Jianguo (Jeff) Xia
University of Alberta, Canada

Metabolomics @ Univ. of Alberta

- To establish infrastructure and to facilitate the application of metabolomics
  - Creating comprehensive, public accessible metabolomic databases
    - The Human Metabolome Database (HMDB)
    - Drugbank, SMPDB
    - MarkerDB, etc
  - Developing robust analytical protocols
    - NMR, GC-MS, LC-MS, DI-MS
    - The Metabolomics Innovation Center (TMIC)
  - Developing bioinformatics tools for metabolomics data analysis

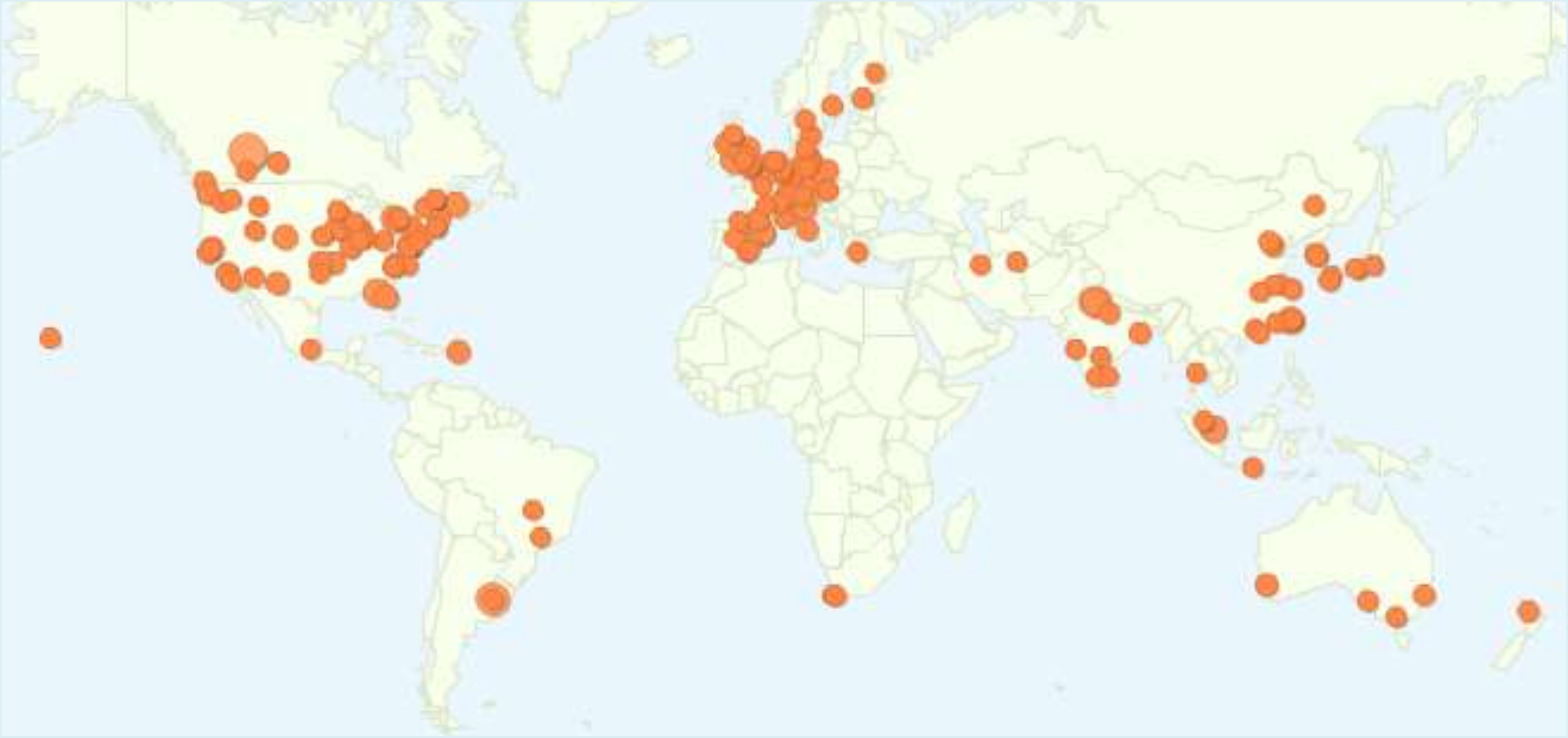
Bioinformatics Tools

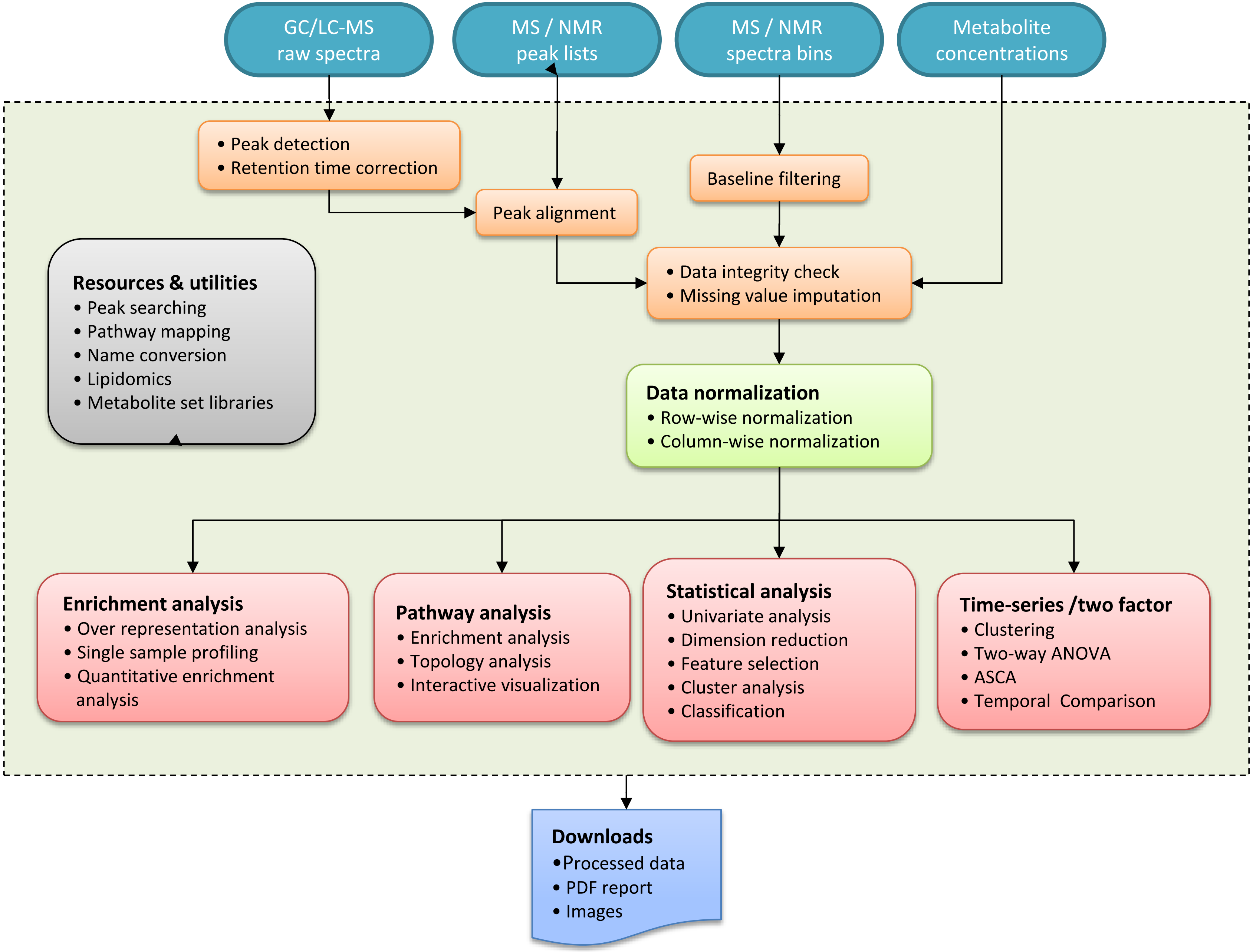
- **MetaboAnalyst**
  - General-purpose metabolomic data analysis pipeline
    - Raw data processing, normalization, statistical analysis …
  - **http://www.metaboanalyst.ca**
- **MSEA** - metabolite set enrichment analysis
  - Identification of enriched groups of metabolites that are biologically meaningful
    - Associated with pathways, diseases, genetic traits ….
  - **http://www.msea.ca**
- **MetPA** - metabolic pathway analysis
  - Identification of metabolic pathways that are most likely to be affected by or involved in the experimental conditions
    - Pathway enrichment analysis, topology analysis, visualization
  - **http://metpa.metabolomics.ca**

❖ *Note, both enrichment analysis and pathway analysis can also be accessed from MetaboAnalyst*

Widely Used

- > 25,000 visits world wide
- ~ 100 jobs /day




GC/LC-MS raw spectra
MS / NMR peak lists
MS / NMR spectra bins
Metabolite concentrations
• Peak detection
• Retention time correction
Peak alignment
Baseline filtering
• Data integrity check
• Missing value imputation
Resources & utilities
• Peak searching
• Pathway mapping
• Name conversion
• Lipidomics
• Metabolite set libraries
Data normalization
• Row-wise normalization
• Column-wise normalization
Enrichment analysis
• Over representation analysis
• Single sample profiling
• Quantitative enrichment analysis
Pathway analysis
• Enrichment analysis
• Topology analysis
• Interactive visualization
Statistical analysis
• Univariate analysis
• Dimension reduction
• Feature selection
• Cluster analysis
• Classification
Time-series /two factor
• Clustering
• Two-way ANOVA
• ASCA
• Temporal Comparison
Downloads
•Processed data
• PDF report
• Images

Motivation

- Metabolomics is very suitable for longitudinal studies
  - Relatively cheaper
  - Convenient sample collection (urine, saliva, blood, etc.)
- Metabolome directly interacts with environment. It is subjected to the effects of various factors
  - Age, sex, life style .....
- ➡ We need bioinformatics tools for time-series and multi-factor metabolomic data analysis

Project Overview

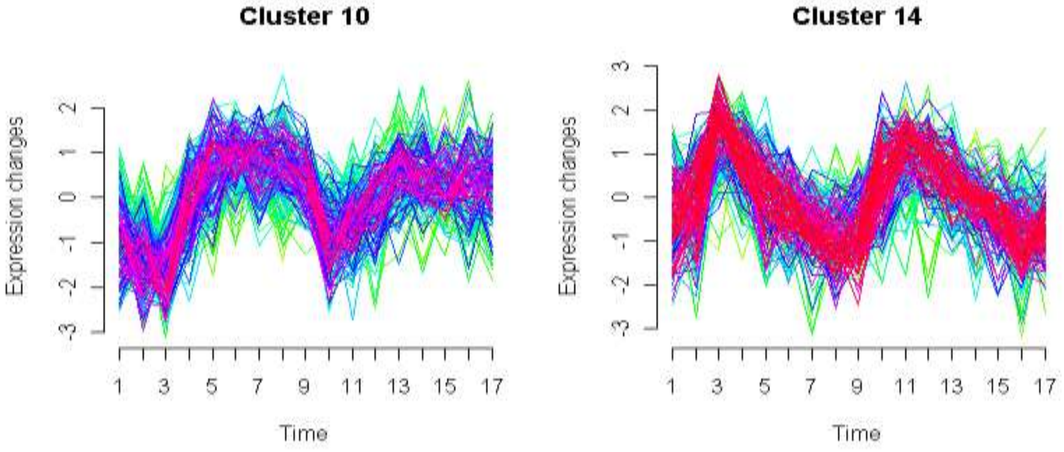
- Objective
  - To develop a user-friendly web-based tool for analyzing time-series / multifactor metabolomic data
- Target users
  - Experimental biologists looking for practical solutions to analyzing their data
- Target experimental design
  - Time and group
  - General two-factor design

Literature Survey

- Common tasks
  - Identification of metabolites with **similar** patterns of change
  - Identification of metabolites with **different** patterns of change
  - Identification of **global patterns of change**
- Commonly used approaches
  - Clustering approaches
  - Linear model (i.e. LIMMA)
  - ANOVA-based approaches (i.e. ASCA)
  - Bayesian approaches (i.e. BATS)
  - Other approaches ….
- Smilde, A.K., *et al.* “Dynamic metabolomic data analysis: a tutorial review” Metabolomics (2010) 6:3–17

Clustering Technique

- Goal: to identify metabolites with similar or correlated expression profiles
- Methods: hierarchical clustering, Kmeans, SOM, etc

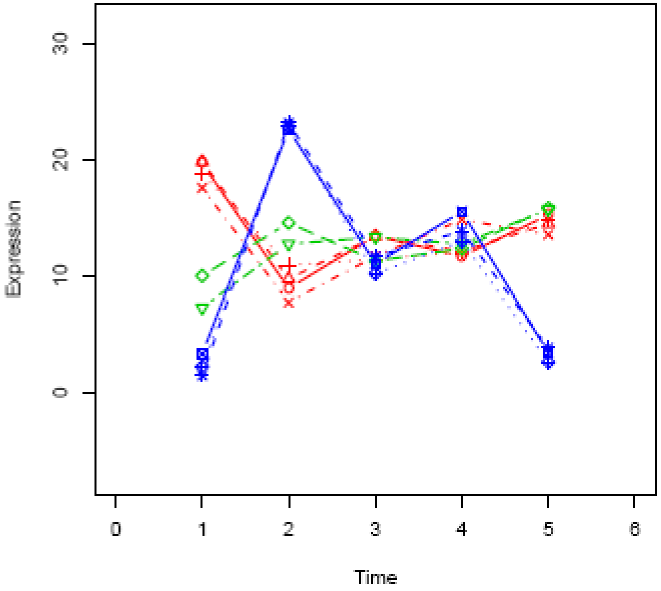


Generated using *Mfuzz* package (L., Kumar and M., Futschik, 2007)

Temporal Profile Analysis

- Goal: to identify temporal profiles that are significantly different among different experimental conditions

- Methods (TimeCourse):
  - Comparing time-course mean profiles of each metabolite
  - Taking consideration of both within and between time points variance



ANOVA-based techniques (ASCA)

- Goal: to identify major trends associated with each experimental factor
- Method: two stages

1. Calculating main effects & interaction

   $$X = A + B + AB + E$$

2. PCA on each partition

   $$X = TP' + E$$

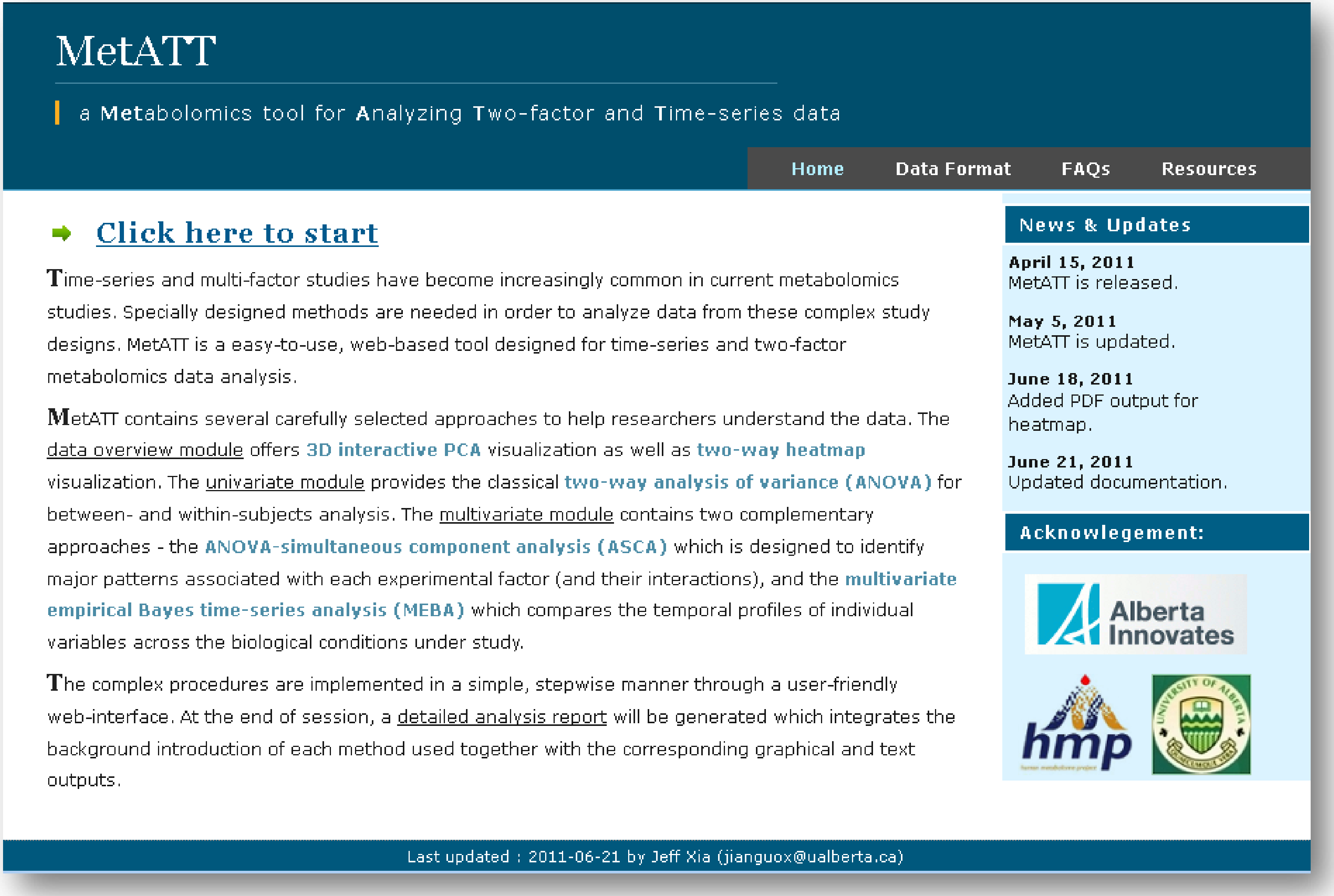
$$X = T_a P'_a + T_b P'_b + T_{ab} P'_{ab} + T_e P'_e + E_a + E_b + E_{ab} + E_e + E$$

Which Methods?

- Selection criteria (for web-based applications)
  - Underlying theory are generally understood (by biologists)
  - Algorithms well established
    - Not proprietary, preferably open-source (i.e. R code)
    - Not computationally prohibitive
  - Relatively straightforward to present

Selected Methods

- Univariate method
  - Two-way within/between-subject ANOVA
- Clustering + visual exploration
  - Interactive 3D PCA
  - Two-way heatmap hierarchical clustering
- Multivariate methods
  - ANOVA – simultaneous component analysis (ASCA)
    - Model selection
    - Model validation
    - Feature selection
  - Multivariate empirical Bayes time-series analysis (MEBA)



MetATT

a Metabolomics tool for Analyzing Two-factor and Time-series data

Click here to start

Time-series and multi-factor studies have become increasingly common in current metabolomics studies. Specially designed methods are needed in order to analyze data from these complex study designs. MetATT is a easy-to-use, web-based tool designed for time-series and two-factor metabolomics data analysis.

MetATT contains several carefully selected approaches to help researchers understand the data. The data overview module offers **3D interactive PCA** visualization as well as **two-way heatmap** visualization. The univariate module provides the classical **two-way analysis of variance (ANOVA)** for between- and within-subjects analysis. The multivariate module contains two complementary approaches - the **ANOVA-simultaneous component analysis (ASCA)** which is designed to identify major patterns associated with each experimental factor (and their interactions), and the **multivariate empirical Bayes time-series analysis (MEBA)** which compares the temporal profiles of individual variables across the biological conditions under study.

The complex procedures are implemented in a simple, stepwise manner through a user-friendly web-interface. At the end of session, a detailed analysis report will be generated which integrates the background introduction of each method used together with the corresponding graphical and text outputs.

News & Updates

April 15, 2011
MetATT is released.

May 5, 2011
MetATT is updated.

June 18, 2011
Added PDF output for heatmap.

June 21, 2011
Updated documentation.

Acknowlegement:

Last updated : 2011-06-21 by Jeff Xia (jianguox@ualberta.ca)

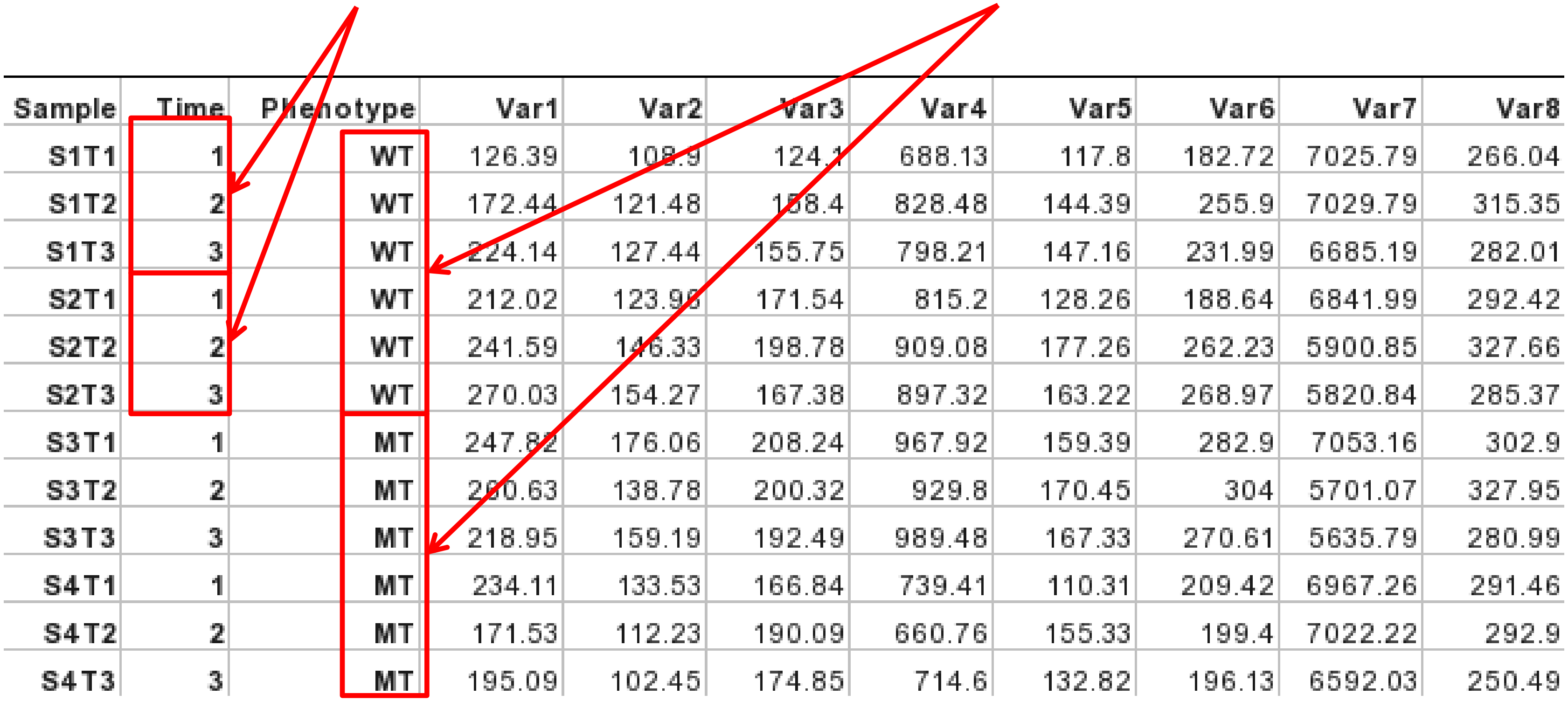
http://metatt.metabolomics.ca

You can also access from within MetaboAnalyst

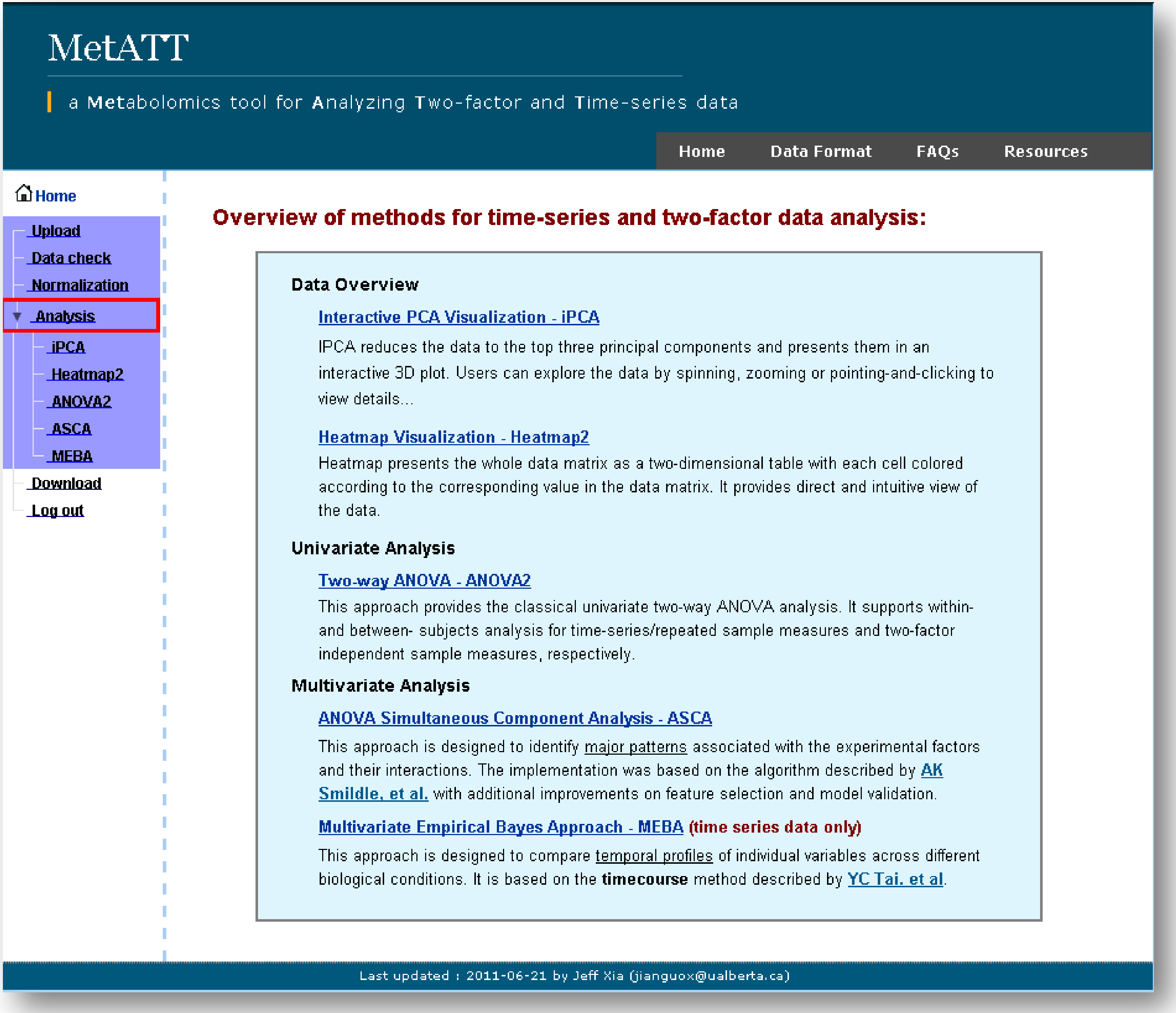
Input Format

Time-series replicates

Experimental conditions



Sample	Time	Phenotype	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8
S1T1	1	WT	126.39	108.9	124.1	688.13	117.8	182.72	7025.79	266.04
S1T2	2	WT	172.44	121.48	158.4	828.48	144.39	255.9	7029.79	315.35
S1T3	3	WT	224.14	127.44	155.75	798.21	147.16	231.99	6685.19	282.01
S2T1	1	WT	212.02	123.96	171.54	815.2	128.26	188.64	6841.99	292.42
S2T2	2	WT	241.59	146.33	198.78	909.08	177.26	262.23	5900.85	327.66
S2T3	3	WT	270.03	154.27	167.38	897.32	163.22	268.97	5820.84	285.37
S3T1	1	MT	247.82	176.06	208.24	967.92	159.39	282.9	7053.16	302.9
S3T2	2	MT	260.63	138.78	200.32	929.8	170.45	304	5701.07	327.95
S3T3	3	MT	218.95	159.19	192.49	989.48	167.33	270.61	5635.79	280.99
S4T1	1	MT	234.11	133.53	166.84	739.41	110.31	209.42	6967.26	291.46
S4T2	2	MT	171.53	112.23	190.09	660.76	155.33	199.4	7022.22	292.9
S4T3	3	MT	195.09	102.45	174.85	714.6	132.82	196.13	6592.03	250.49



MetATT

a Metabolomics tool for Analyzing Two-factor and Time-series data

Overview of methods for time-series and two-factor data analysis:

Data Overview

Interactive PCA Visualization - iPCA

IPCA reduces the data to the top three principal components and presents them in an interactive 3D plot. Users can explore the data by spinning, zooming or pointing-and-clicking to view details...

Heatmap Visualization - Heatmap2

Heatmap presents the whole data matrix as a two-dimensional table with each cell colored according to the corresponding value in the data matrix. It provides direct and intuitive view of the data.

Univariate Analysis

Two-way ANOVA - ANOVA2

This approach provides the classical univariate two-way ANOVA analysis. It supports within- and between- subjects analysis for time-series/repeated sample measures and two-factor independent sample measures, respectively.

Multivariate Analysis

ANOVA Simultaneous Component Analysis - ASCA

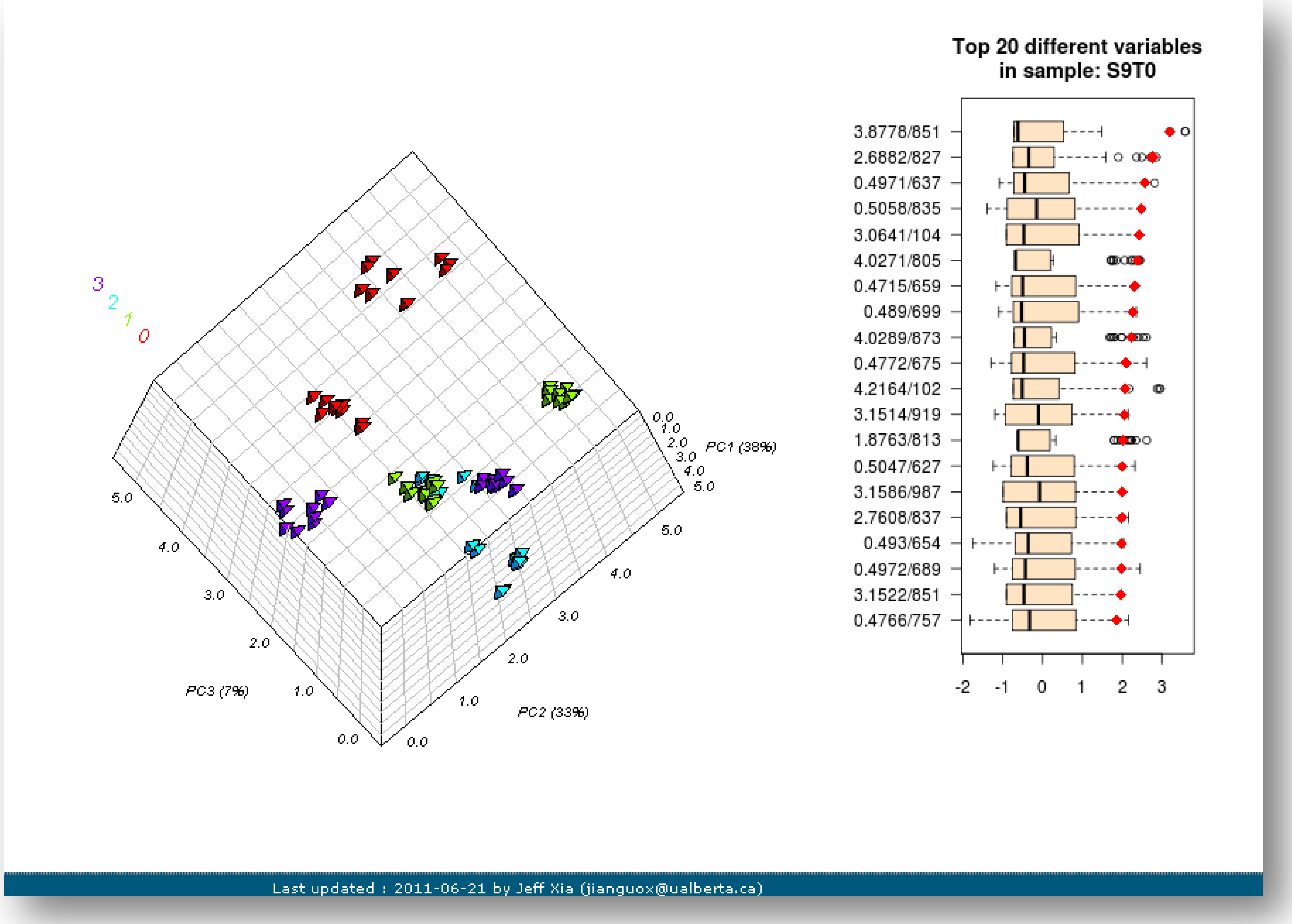
This approach is designed to identify major patterns associated with the experimental factors and their interactions. The implementation was based on the algorithm described by AK Smildle, et al. with additional improvements on feature selection and model validation.

Multivariate Empirical Bayes Approach - MEBA (time series data only)

This approach is designed to compare temporal profiles of individual variables across different biological conditions. It is based on the **timecourse** method described by YC Tai, et al.

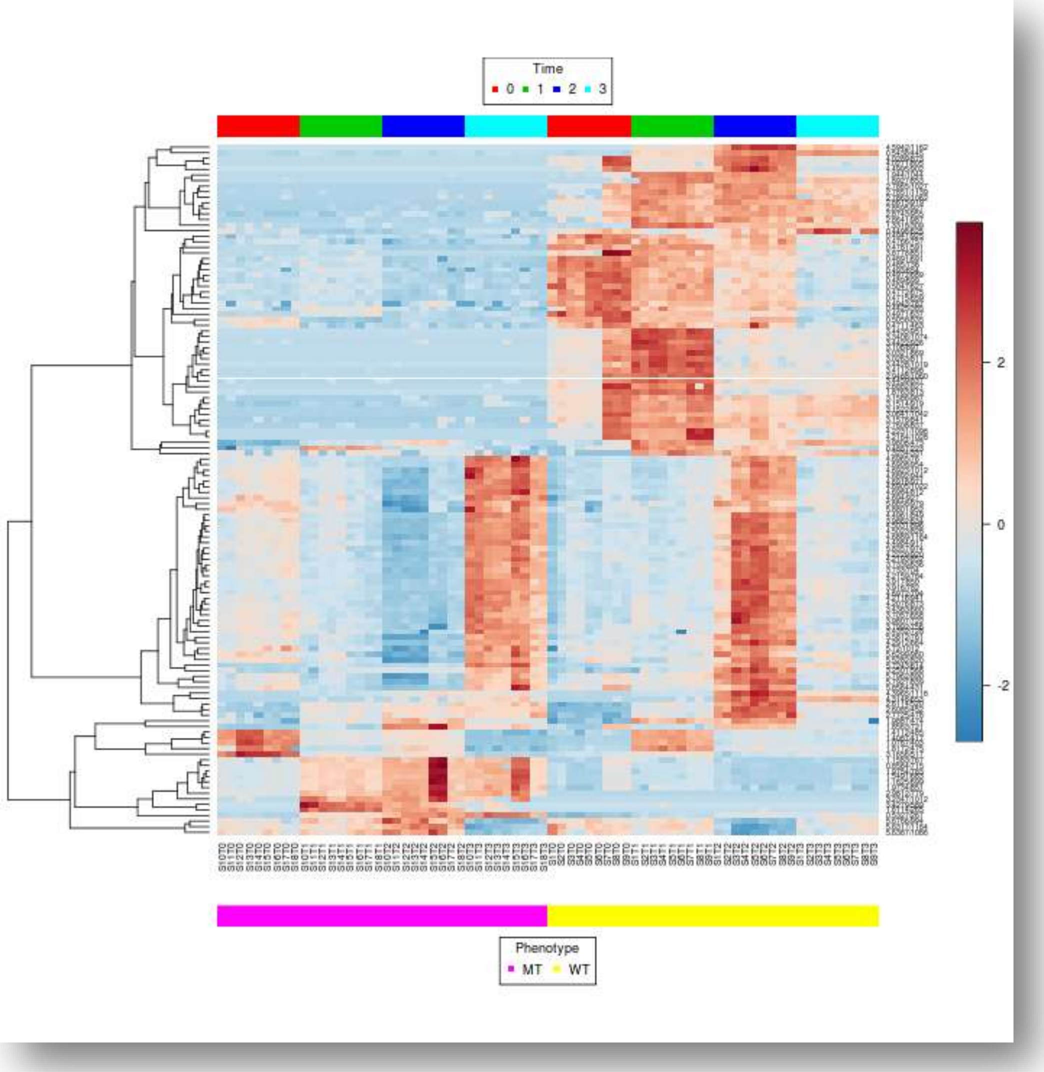
Last updated : 2011-06-21 by Jeff Xia (jianguox@ualberta.ca)

Interactive 3D PCA visualization

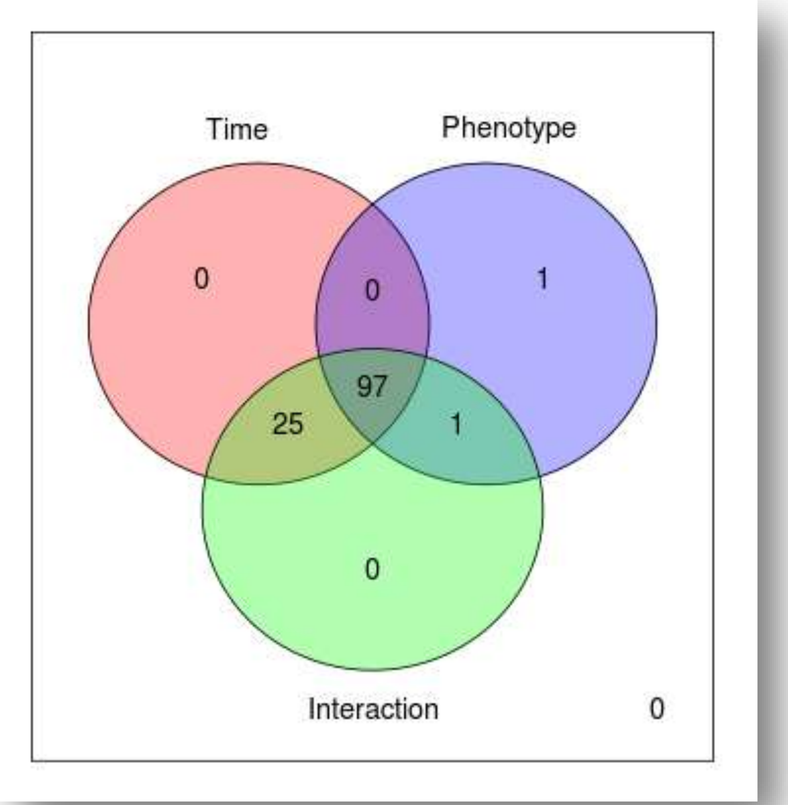


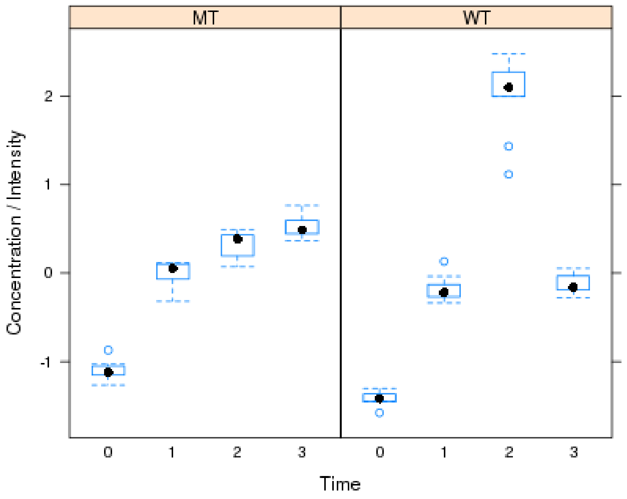
Last updated : 2011-06-21 by Jeff Xia (jianguox@ualberta.ca)

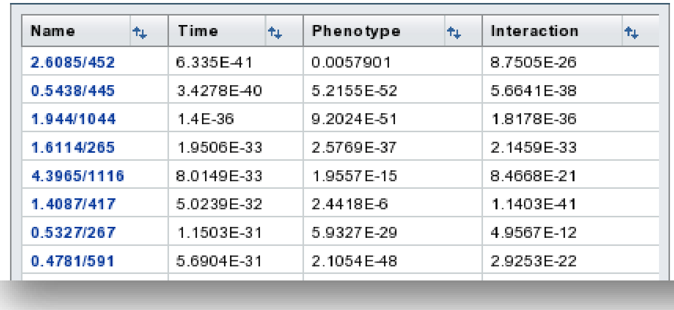
Two-way clustered heatmap



Two-way ANOVA

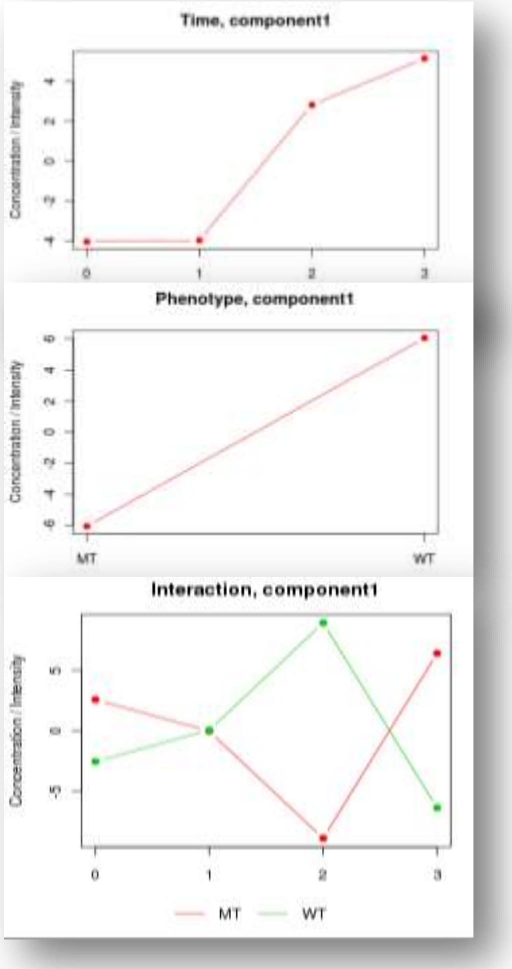


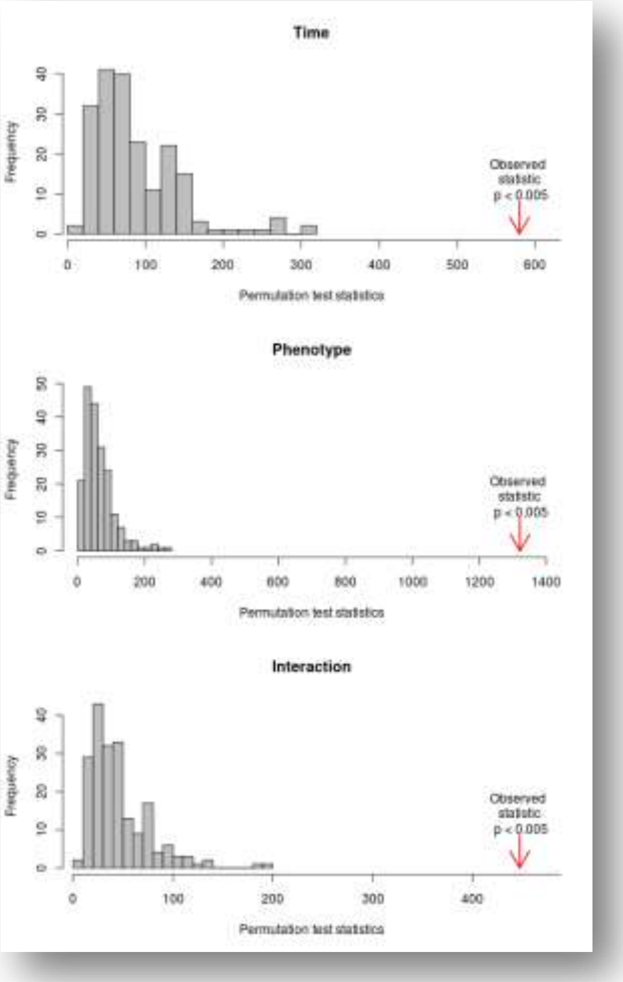




Name	Time	Phenotype	Interaction
2.6085/452	6.335E-41	0.0057901	8.7505E-26
0.5438/445	3.4278E-40	5.2155E-52	5.6641E-38
1.944/1044	1.4E-36	9.2024E-51	1.8178E-36
1.6114/265	1.9506E-33	2.5769E-37	2.1459E-33
4.3965/1116	8.0149E-33	1.9557E-15	8.4668E-21
1.4087/417	5.0239E-32	2.4418E-6	1.1403E-41
0.5327/267	1.1503E-31	5.9327E-29	4.9567E-12
0.4781/591	5.6904E-31	2.1054E-48	2.9253E-22

ASCA – identification of major trends

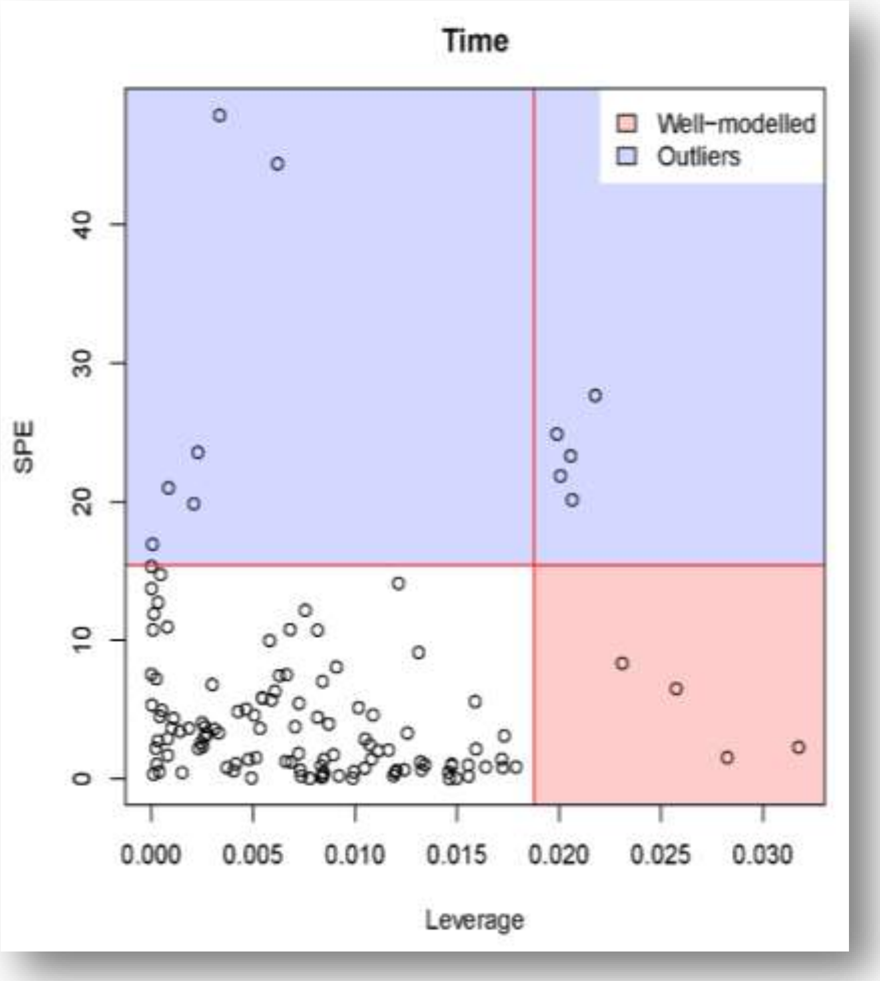




Smilde, A.K., *et al.* and Vis, D.J., *et al.*

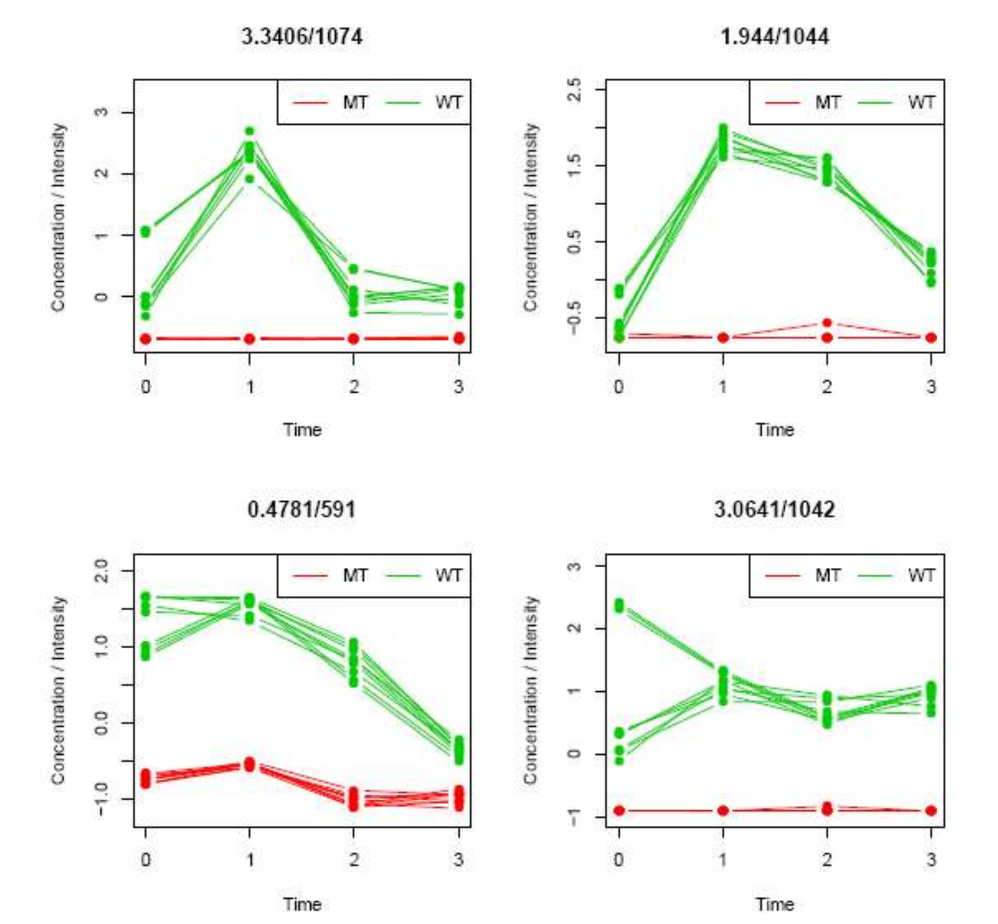
ASCA – variable selection

- To identify significant variables that behave according to the detected patterns.
  - Leverage
    - Derived from loadings
  - Squared prediction error (SPE)
    - Derived from residuals
- Nueda, et al., 2007

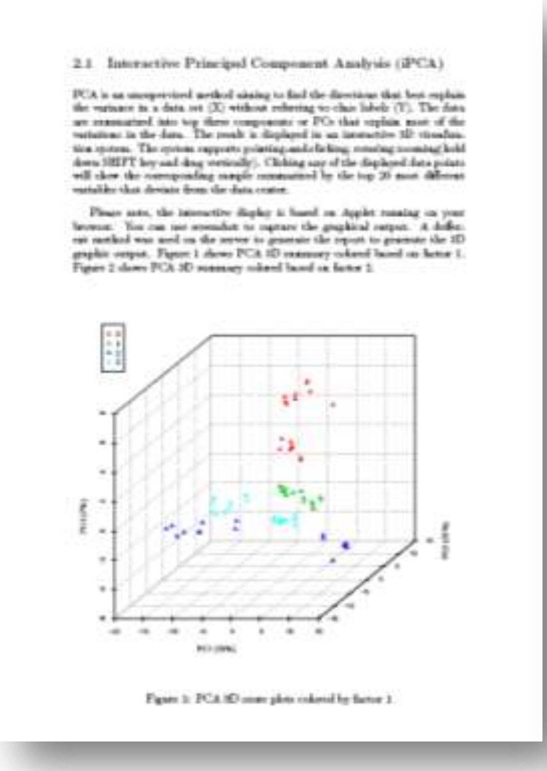


Temporal Profile Analysis

- To detect variables with distinctive temporal profiles
- The variables are ranked by *MB-statistics*
- Tai, Y.C. and Speed, T.P.



Analysis Report (PDF)

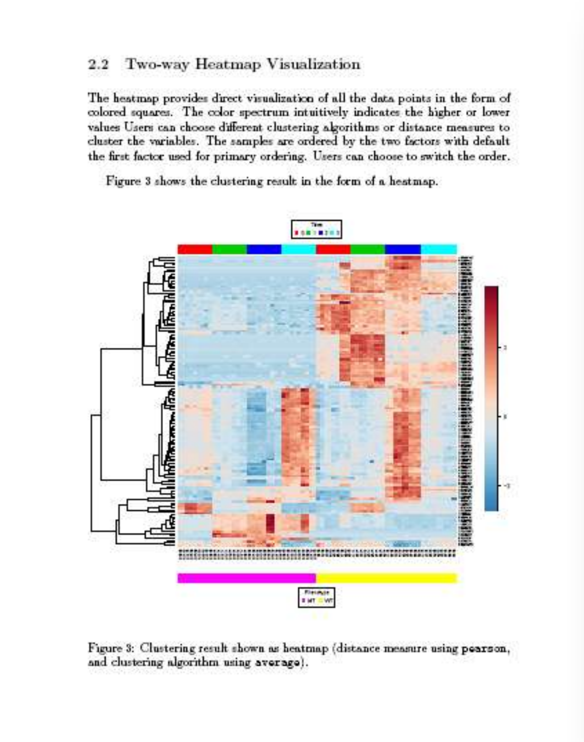


2.1 Interactive Principal Component Analysis (iPCA)

PCA is an unsupervised method aiming to find the directions that best explain the variance in a data set (X) without referring to class labels (Y). The data are summarized into top three components or PCs that explain most of the variations in the data. The result is displayed in an interactive 3D visualization system. The system supports pointing-and-clicking, rotating, zooming(hold down SHIFT key and drag vertically). Clicking any of the displayed data points will show the corresponding sample summarized by the top 20 most different variables that deviate from the data center.

Please note, the interactive display is based on Applet running on your browser. You can use screenshot to capture the graphical output. A different method was used on the server to generate the report to generate the 3D graphic output. Figure 1 shows PCA 3D summary colored based on factor 1. Figure 2 shows PCA 3D summary colored based on factor 2.

Figure 1: PCA 3D score plots colored by factor 1

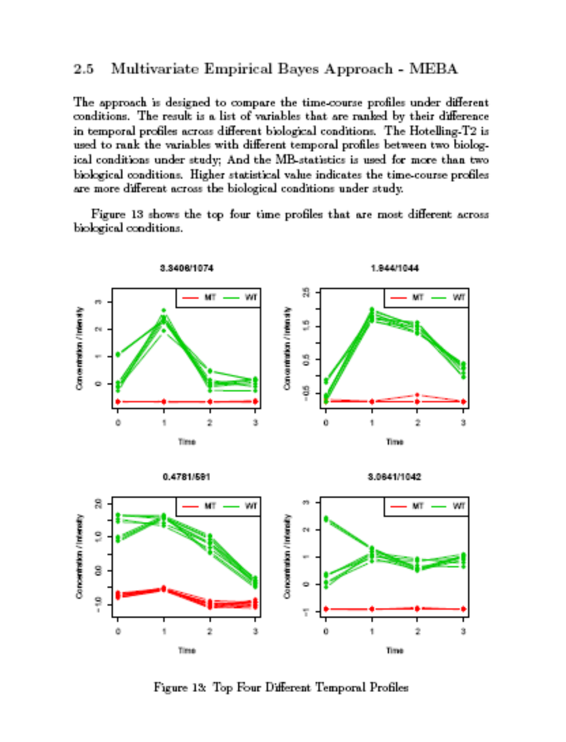


2.2 Two-way Heatmap Visualization

The heatmap provides direct visualization of all the data points in the form of colored squares. The color spectrum intuitively indicates the higher or lower values Users can choose different clustering algorithms or distance measures to cluster the variables. The samples are ordered by the two factors with default the first factor used for primary ordering. Users can choose to switch the order.

Figure 3 shows the clustering result in the form of a heatmap.

Figure 3: Clustering result shown as heatmap (distance measure using pearson, and clustering algorithm using average).



2.5 Multivariate Empirical Bayes Approach - MEBA

The approach is designed to compare the time-course profiles under different conditions. The result is a list of variables that are ranked by their difference in temporal profiles across different biological conditions. The Hotelling-T2 is used to rank the variables with different temporal profiles between two biological conditions under study; And the MB-statistics is used for more than two biological conditions. Higher statistical value indicates the time-course profiles are more different across the biological conditions under study.

Figure 13 shows the top four time profiles that are most different across biological conditions.

Figure 13: Top Four Different Temporal Profiles

Tips

- Data should be centered (zero-means) and properly normalized before analysis
- When there appear to be a strong interaction effect, the result need to interpreted in the context of each other; otherwise, it is advisable to analyzed the data with regard to each factor independently for ease of interpretation
- Multi (> 2) factor data analysis can be decomposed into two-factor data analysis

Summary

- MetATT offers three commonly used approaches
  - i. Identification of compounds with similar patterns - PCA, heatmap
  - ii. Identification of compounds with different temporal profile analysis - MEBA, ANOVA
  - iii. Identification of global major (and secondary) patterns - ASCA
- Time-series and multi-factor omics data analysis is still at its early stages.
  - No well-defined procedures
- Future efforts
  - Evaluate other approaches
    - PARAFAC, Tucker, N-PLS, etc
  - Challenges
    - Too complex to be presented on a web-based program
    - Lack of widely-accepted outputs & associated interpretations

References

- 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, *Annals of Statistics*, **34**, 2387-2412.
- Vis, D.J., *et al.* (2007) Statistical validation of megavariate effects in ASCA, *BMC Bioinformatics*, **8**, 322.
- Nueda, M.J., *et al.* (2007) Discovering gene expression patterns in time course microarray experiments by ANOVA-SCA, *Bioinformatics*, **23**, 1792-1800.
- Glaab, E., *et al.* (2010) vrmlgen: An R Package for 3D Data Visualization on the Web, *Journal of Statistical Software*, **36**, 2347-2348.

Acknowledgements

- Dr. David Wishart
- Dr. Rupa Mandal
- Dr. Igor Sinelnikov
- Dr. David Broadhurst

R

Java

Alberta Innovates Technology Futures

University of Alberta
Quaecumque Vera

hmp
human metabolome project