metabox: a toolbox for metabolomic data analysis, interpretation and integrative exploration

Kwanjeera Wanichthanarak, Sili Fan, Dmitry Grapov, Dinesh Barupal and Oliver Fiehn

9 January 2017

TUTORIALS

A collection of tutorials for using metabox is provided here to illustrate important features of metabox. These tutorials are corresponding to Results and Discussion section of the main manuscript.

0. Example data sets

The example data sets are based on the published studies about lung adenocarcinoma (transcriptomic data (1) and metabolomic data (2)).

The metabolomic data set containing 39 malignant and adjacent non-malignant lung tissue samples was measured by gas chromatography time-of-flight mass spectrometry (GC-TOF-MS) (2) and pre-processed by the BinBase database (3). 462 compounds were measured, 171 of which have associated PubChem CID. This metabolomic data set was prepared in the input format used in metabox, which is in metabolomics.xlsx.

Metabox was used for the statistical analysis comparing tumor and control tissues. For users’ convenience, we filtered the list of significantly different compounds comparing tumor and control tissues from the analysis result and provided in pubchem_stats.txt.

The transcriptomic data, GEO accession: GSE32863 (1), contain gene expression profiles of 58 lung adenocarcinoma and adjacent non-malignant lung tissues using Illumina HumanWG-6 v3.0 expression BeadChip platform. Differential gene expression analysis comparing lung adenocarcinoma and adjacent non-malignant lung tissues was performed with GEO2R (4). P-values were adjusted with Benjamini and Hochberg false discovery rate at 5% (pFDR < 0.05). Differential gene expression analysis comparing tumor and control tissues reported 171 out of 21,204 genes in which pFDR < 0.05 and |log2FC| > 2. For users’ convenience, we filtered the list of significantly different genes comparing tumor and control tissues from the analysis result and provided in ensembl_stats.txt.

Table 1. List of example data sets

<table>
<thead>
<tr>
<th>File</th>
<th>Description</th>
<th>Input of section</th>
</tr>
</thead>
<tbody>
<tr>
<td>metabolomics.xlsx</td>
<td>Metabolic profiles of 39 lung malignant and adjacent non-malignant tissues pre-processed by the BinBase database.</td>
<td>Section 2</td>
</tr>
<tr>
<td>pubchem_stats.txt</td>
<td>Multi-column table of significant compounds and associated statistical values computed by metabox.</td>
<td>Section 3</td>
</tr>
<tr>
<td>ensembl_stats.txt</td>
<td>Multi-column table of filtered genes and associated statistical values computed by GEO2R.</td>
<td>Section 3</td>
</tr>
</tbody>
</table>
1. Load required libraries and run the metabox using R-terminal

```r
## Load library
> library(metabox)
> library(opencpu)
```

```r
## Run metabox on a web browser
> opencpu$browse('library/metabox/www')
```

2. Using metabox for in-depth analysis of metabolomic data

This tutorial shows that metabox is used for deep analysis (data processing, statistical analysis, metabolic network construction and functional interpretation) of metabolomic data. The example metabolomic data set is in `metabolomics.xlsx`. It is uploaded to metabox for log transformation before statistical analysis using paired t-test (Fig 1). The output from the data normalization and the statistical analysis is transferred to the network construction part to calculate the chemical structure similarity network of the PubChem compounds. The default threshold of correlation coefficient > 0.7 was used. The resulting network is enhanced by further mapping with annotation information. In this case, we applied Functional class scoring option to estimate significantly enriched pathways of network nodes by taking metabolic profiles in to account (see Fig 2 for all the steps).

Alternatively, the output from statistical analysis can be passed to one of the functional analysis method to aid interpretation of measured compounds. In this case, we applied WordCloud option to quickly summarize KEGG pathways of the compounds (see Fig 3 for all the steps). The top ten categories of measured compounds include Central carbon metabolism in cancer, Protein digestion and absorption, Aminoacyl-tRNA biosynthesis, Purine metabolism, Pyrimidine metabolism, Glyoxylate and dicarboxylate metabolism, Arginine and proline metabolism, Pentose and glucuronate interconversions, Galactose metabolism and Fatty acid biosynthesis sorted by the number of compounds in each category. This example shows that after statistical analyses, the result can be interpreted with pathway information.

3. Integrative exploration of significant genes and compounds in biological network context

Metabox is used for joint exploration of the lists of significant genes and compounds from comparisons between tumor and non-tumor tissues in the context of biological networks. The `ensemble_stats.txt` and `pubchem_stats.txt` are used to construct the biological network outlining relationships between genes and compounds as `(Protein)-[:CONTROL]->(from:Gene)-[:CONVERSION]->(Protein)-[:CATALYSIS]->(to:Compound)`. The resulting network including all attributes can be downloaded for advance exploration using visualization software such as Cytoscape (5). Fig 4 shows all the steps to construct the network with metabox.
Fig 1. Data normalization and statistical analysis comparing two-paired groups of metabolomic study.
Fig 2. Computing chemical similarity network and performing FnClassScoring on network nodes.
3. Choose Pathway and Click FinClass/Scoring

Functional class scoring

- Estimate enriched functional classes for the input.

| Entity link (e.g., PubChem or enzyme or annotated) will be used for enrichment analysis.

<table>
<thead>
<tr>
<th>PathChem</th>
<th>entity</th>
<th>p-value</th>
<th>adj p-value</th>
<th>non-parametric p-value</th>
<th>non-parametric adj p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0012</td>
<td>term</td>
<td>0.0012</td>
<td>0.0022</td>
<td>0.0022</td>
<td>0.0022</td>
</tr>
<tr>
<td>09459</td>
<td>true</td>
<td>0.0286</td>
<td>0.0548</td>
<td>0.0548</td>
<td>0.0548</td>
</tr>
<tr>
<td>02195</td>
<td>true</td>
<td>0.0021</td>
<td>0.0071</td>
<td>0.0071</td>
<td>0.0071</td>
</tr>
<tr>
<td>0607</td>
<td>false</td>
<td>0.0844</td>
<td>0.0719</td>
<td>0.0719</td>
<td>0.0719</td>
</tr>
<tr>
<td>00005</td>
<td>true</td>
<td>0.0845</td>
<td>0.0564</td>
<td>0.0564</td>
<td>0.0564</td>
</tr>
<tr>
<td>02009</td>
<td>true</td>
<td>0.0892</td>
<td>0.0815</td>
<td>0.0815</td>
<td>0.0815</td>
</tr>
<tr>
<td>0175</td>
<td>true</td>
<td>0.0888</td>
<td>0.0888</td>
<td>0.0888</td>
<td>0.0888</td>
</tr>
<tr>
<td>0176</td>
<td>true</td>
<td>0.0840</td>
<td>0.0840</td>
<td>0.0840</td>
<td>0.0840</td>
</tr>
<tr>
<td>0174</td>
<td>true</td>
<td>0.0844</td>
<td>0.0844</td>
<td>0.0844</td>
<td>0.0844</td>
</tr>
<tr>
<td>0173</td>
<td>false</td>
<td>0.0844</td>
<td>0.0844</td>
<td>0.0844</td>
<td>0.0844</td>
</tr>
</tbody>
</table>

Result of statistical analysis is shown in interactive tables. PubChem column and statistical values will be used to compute enriched functional classes.

4. Choose method median, Select Pathway and Select column p-value, Non-adj. and Click Compute.

Table of analysis result, node attributes, edge lists and annotation-entity pairs are shown. Top ten terms are colored. Colors in the table and nodes are the same. Color legend is in the Network panel.

The resulting network will be mapped with annotation terms. Nodes that are not the parts of the top ten annotations or not annotated are in grey.

Click Download Functional analysis results will download results of functional analysis and the network mapped with annotation terms including corresponding legend.
Fig 3. Computing WordCloud for measured compounds.
Fig 4. Constructing biological network from gene lists and compound lists. Relationship pattern between genes and compounds is:

\[ (:Protein) \rightarrow [:CONTROL] \rightarrow (:Gene) \rightarrow [:CONVERSION] \rightarrow (:Protein) \rightarrow [:CATALYSIS] \rightarrow (to:Compound) \]
REFERENCES


