

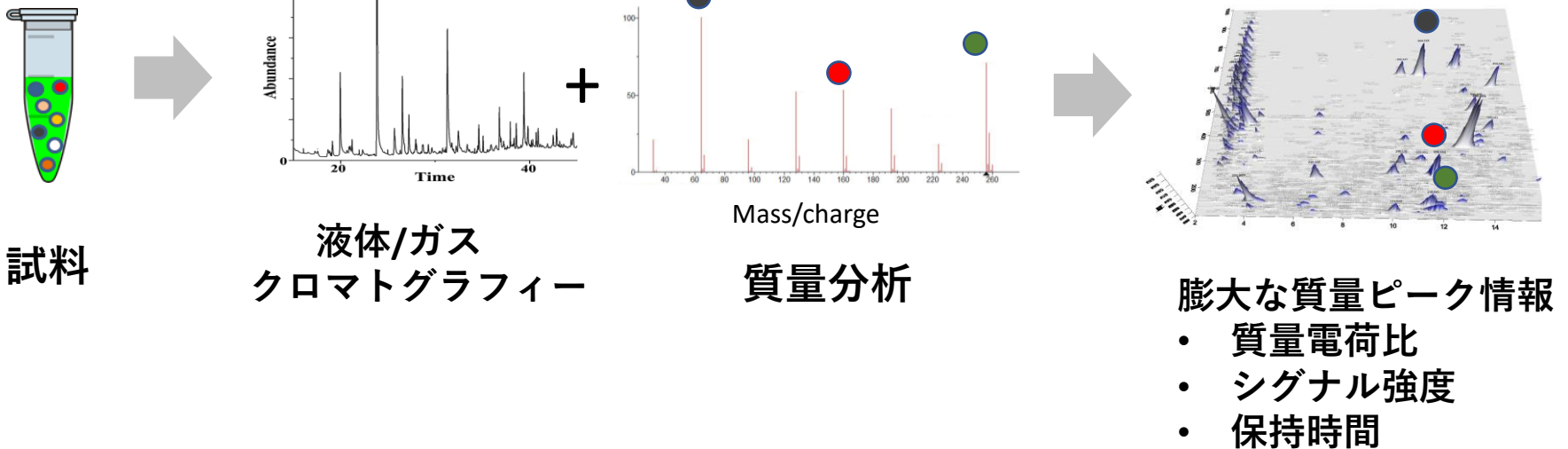
# メタボロームデータ解析と活用法： データベースとオンラインツールの使い方

早川 英介

理化学研究所 環境資源科学研究センター  
メタボローム情報研究チーム

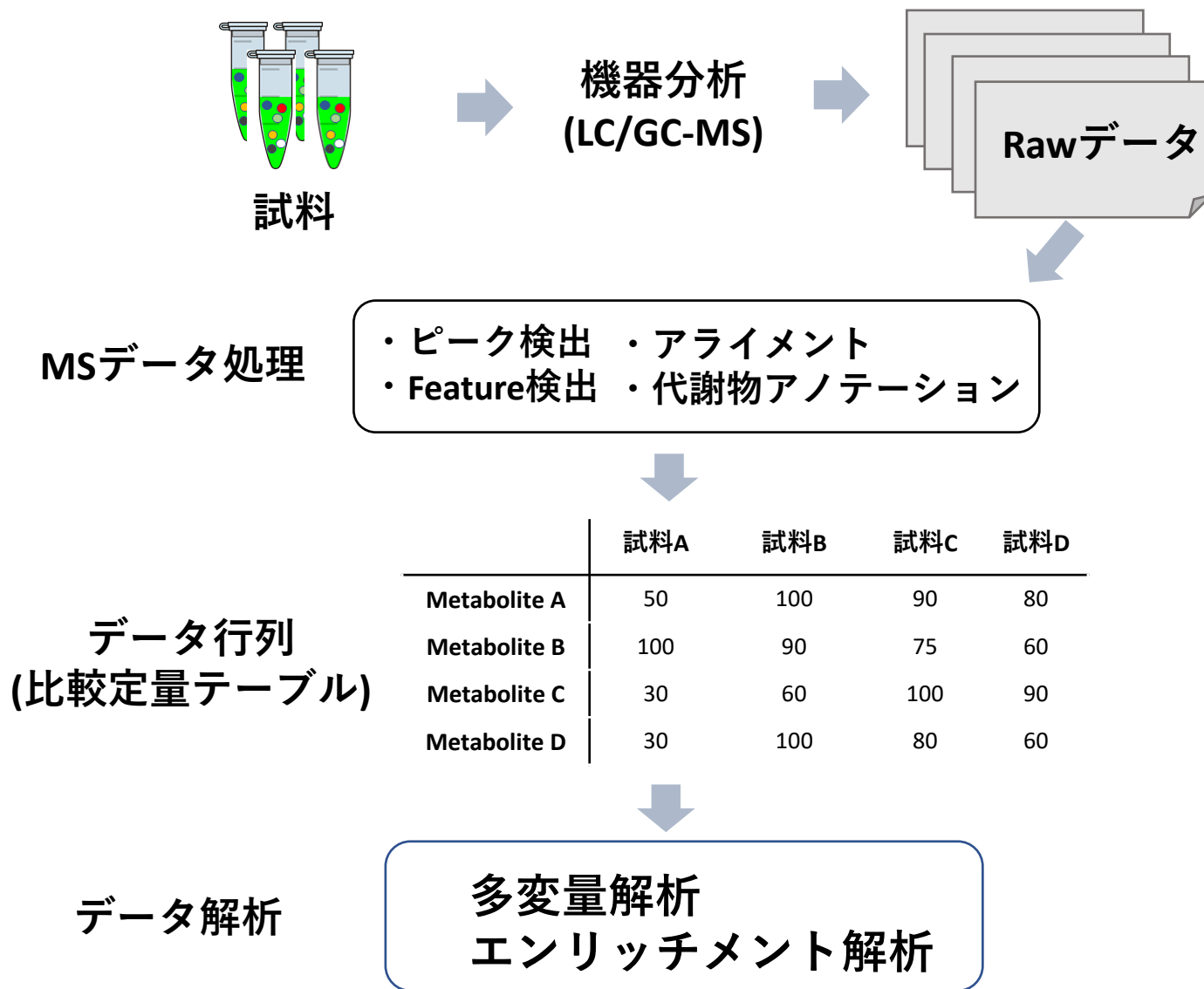
- **本講義の目的**
  - メタボロミクスデータの解析法を理解し、研究に活用する
- **本講義の対象**
  - 質量分析のデータ解析に慣れていない方
  - 他のオミクス研究者
- **講義内容**
  - メタボロミクスデータの簡単な解説
  - 公共レポジトリ
  - データ解析ツール

# 質量分析によるメタボローム分析

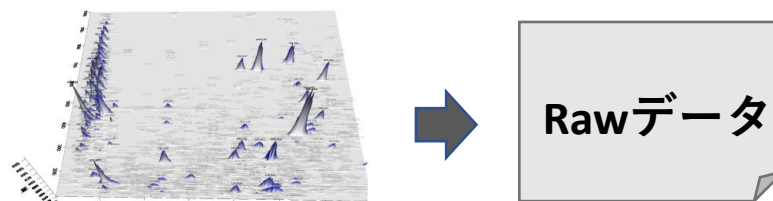


- ターゲット分析：  
あらかじめ測定対象（ターゲット）を決めて分析を行う
- ノンターゲット分析：  
測定対象を設定せずに分析を行い、検出された質量ピーク全てに関して解析を行う

# メタボロミクスデータ処理フロー



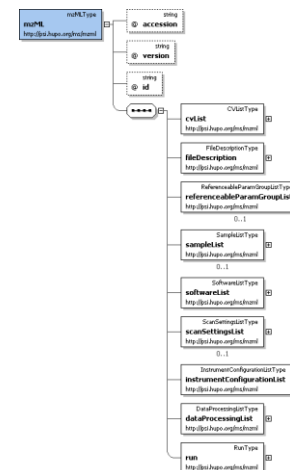
# 質量分析におけるRAWデータ



- 生・未加工のLC/GC-MSデータ
- 装置のセッティング・質量電荷比・ion count値の配列・クロマトグラム情報
- ベンダーが提供しているSDK等で中身へのアクセスは可能

## 公共フォーマットmzML


- Open format
- XMLベース
- ほとんどのMS関連ソフト・ツールで読み込み可
- Msconvert (ProteoWizard) でベンダー形式から変換可能



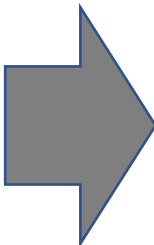
# RAW データの中身

## .mzML

- 装置
- 測定条件
- その他の情報
  - イメージング : XY位置情報
  - イオンモビリティ : Drift bin

- 
- Scan number
  - spectrum type (MS1 or 2)
  - m/z 配列
  - Intensity 配列

## MSデータ処理

- 
- スペクトル処理
  - データアライメント
  - Mass feature検出
  - データ行列生成

# メタボロミクスのデータレポジトリ



## **Metabolights (EMBL-EBI)**

メタボロミクス実験と関連情報のレポジトリ (1333 studies)



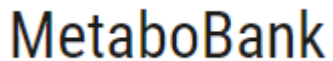
## **Metabolomics Workbench (UC San Diego)**

メタボロミクスデータベース・レポジトリ・解析ツール  
(2747 studies)



## **Metabolomic Repository Bordeaux**

植物に特化したメタボロミクスレポジトリ (58 projects)



## **MetaboBank (DDBJ)**

メタボロミクスデータレポジトリ。MAGE-TAB formatによるメタデータの記述(104 studies)

- Rawデータに加え試料・実験条件などのメタデータを格納
- 結果の再現性・データ共有の観点から論文投稿時に強く求められる
- 再解析や他のオミクスデータとの連携など利用価値は高い

# レポジトリデータの活用法・再解析

- 興味ある現象・試料
- 他オミクス (geno-/transcript/proteo)



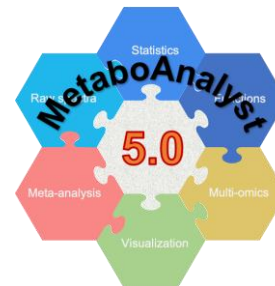
## メタボロミクスレポジトリ



- 実験情報
- Rawデータ
- データ行列



## データ解析プラットフォーム



- 多変量解析
- エンリッチメント解析



- 代謝パスウェイ情報
- 代謝物情報
- 生物学的解釈

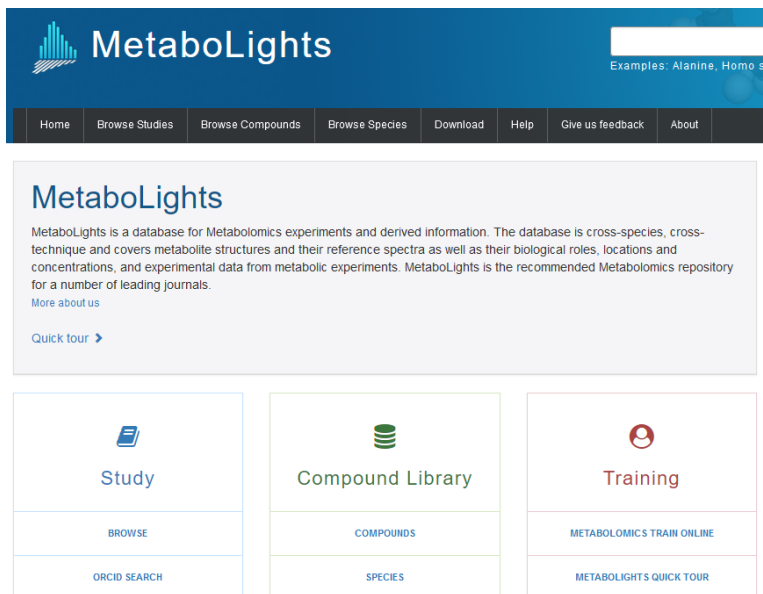


- 代謝物の観点からの知見
- オミクスデータ統合



# Metabolights

- メタボロミクス研究のためのデータベース/レポジトリ
- Rawデータ、実験情報、データ行列（比較定量テーブル）等が登録
- Cross-species, cross-technique
- 化合物情報など関連情報、外部データベースとの接続性



- Studies: プロジェクト単位での各種データのレポジトリ
- Compounds: 代謝物構造データベース

# レポジトリ上のデータのブラウジング (Browse Studies)

The screenshot shows the MetaboLights website interface. At the top is a navigation bar with links: Home, Browse Studies, Browse Compounds, Browse Species, Download, Help, Give us feedback, About, Submit Study, and Login. Below the navigation bar is a search bar with the text 'MetaboLights / Search'. On the left side, there are three filter panels. The first panel, 'Filter your results', has a 'Type' dropdown set to 'study' and a 'Technology' dropdown set to 'Mass spectrometry'. The second panel, 'Organism', has a search bar and a list of organisms with checkboxes, including 'Homo sapiens', 'Mus musculus', 'Rattus norvegicus', 'Saccharomyces cerevisiae', 'Escherichia coli', 'Vitis vinifera', 'reference compound', 'Caenorhabditis elegans', 'Arabidopsis thaliana', 'Daphnia magna', 'Chaetomium globosum', 'Streptomyces species', and 'Mycoplasma genitalium'. The third panel, 'Organism Part', has a search bar and a list of organism parts with checkboxes, including 'blood', 'blood serum', 'Whole Organism', 'Liver', 'Urine', 'Liver', and 'Whole organism'. The main content area shows '1322 results, showing 1 to 10'. A red box highlights the 'ヒットしたStudies' (Hit Studies) section, which contains a table of search results. The table has columns for 'Study Identifier', 'Study Size', 'Submitted by', 'Organism', and 'Study Factors'. The first result is 'MTBL31324' by 'Dylan Zeiss' for 'Organism: Mus musculus'. The second result is 'MTBL33505' by 'Ilu kun' for 'Organism: Homo sapiens'. The title of the second study is 'UPLC-Q-TOF/MS-Based Plasma and Urine Metabolomics Contribute to the Diagnosis of Sepsis.'

ヒットしたStudies

Study Identifier	Study Size	Submitted by	Organism	Study Factors
MTBL31324	34.48GB	Dylan Zeiss	Mus musculus	Treatment, Timepoint, Biological replicate
MTBL33505	42.29GB	Ilu kun	Homo sapiens	Disease

- Metabolights上には多数の"Study"として研究データセットがRawデータ、定量テーブル、各種メタデータと共に格納されている。  
(1333 studies 2023Aug)

- Studyを
  - 分析法：Mass spectrometry/NMR
  - Organism
  - Organism partで検索して取得することが可能

<https://www.ebi.ac.uk/metabolights/>

# 各Studyのデータ

Status Public Release Date 2023-04-18

## MTBLS3505: UPLC-Q-TOF/MS-Based Plasma and Urine Metabolomics Contribute to the Diagnosis of Sepsis.

Su Guan, Kun Liu, Zimeng Liu, Liping Zhou, Bingjie Jia, Zichen Wang, Yao Nie, Xuyu Zhang

In this study, we aimed to identify potential metabolic biomarkers that can improve the diagnostic accuracy of sepsis. Sixty-six patients including 30 septic and 36 nonsepsis patients from an intensive care unit were recruited. The global plasma and urine metabolomic profiles were determined by ultraperformance liquid chromatography coupled with a quadrupole time-of-flight mass spectrometry-based methodology. The risk factors, including both traditional physiological indicators and metabolic biomarkers, were investigated by binary logistic regression analysis and used to build a least absolute shrinkage and selection operator (Lasso) regression model to evaluate the ability of diagnosis. Fifty-five metabolites in plasma and 11 metabolites in urine were identified through orthogonal projections to latent structures discriminant analysis (OPLS-DA). Among them, ten (PE (20:4(5Z, 8Z, 11Z, 14Z)/P-18:0), harderoporpyrinogen, chloropanaxydiol, (Z)-2-octenal,  $N^1,N^8$ -diacetylspermidine, 1-nitroheptane, venoterpine,  $\alpha$ -CEHC, LysoPE (20:0/0:0), corticocin) metabolites were identified as risk factors. The Lasso regression model incorporating these ten metabolic biomarkers and five traditional physiological indicators displayed better differentiation than the traditional model, represented by the elevated area under receiver operating characteristic curve (AUROC) from 96.80 to 100.0%. Furthermore, patients with septic shock presented a significantly lower level of PE-Cer (d16:1(4E)/19:0). This study suggests that metabolomic profiling could be an effective tool for sepsis diagnosis.

### PUBLICATIONS

#### UPLC-Q-TOF/MS-Based Plasma and Urine Metabolomics Contribute to the Diagnosis of Sepsis.

44. Guan S, Liu K, Liu Z, Zhou L, Jia B, Wang Z, Nie Y, Zhang X.

Descriptors Protocols Samples Assays Metabolites Files

#### KEYWORDS

Untargeted Metabolites Ultra-Performance Liquid Chromatography-Mass Spectrometry Diagnosis Sepsis, CTCAE

#### Factors

Disease : Disease

- **Descriptors:**  
studyの内容を示すkeyword
- **Protocols:**  
試料調整・MS分析法・アノテーション法まで含む処理の概要
- **Samples:**  
各サンプルの概要
- **Assays:**  
各サンプルで行われた一連の処理の詳細
- **Metabolites:**  
Studyで検出された代謝物のリスト
- **Files:**  
各種ファイル

<https://www.ebi.ac.uk/metabolights/editor/MTBLS3505>





















# ファイル構成

Descriptores   Protocols   Samples   Assays   Metabolites   **Files**

FTP Download   Aspera Download   ? Help

☐ ISA METADATA Download

search meta data

<input type="checkbox"/>	 a_MTBLS3505_plasma_lip_neg_metabolite_profiling_mass_spectrometry.txt	April 18 2023 17:39:08	
<input type="checkbox"/>	 a_MTBLS3505_plasma_lip_pos_metabolite_profiling_mass_spectrometry.txt	April 18 2023 17:39:07	
<input type="checkbox"/>	 a_MTBLS3505_plasma_met_neg_metabolite_profiling_mass_spectrometry.txt	April 18 2023 17:39:07	
<input type="checkbox"/>	 a_MTBLS3505_plasma_met_pos_metabolite_profiling_mass_spectrometry.txt	April 18 2023 17:39:07	
<input type="checkbox"/>	 a_MTBLS3505_urine_met_neg_metabolite_profiling_mass_spectrometry.txt	April 18 2023 17:39:07	
<input type="checkbox"/>	 a_MTBLS3505_urine_met_pos_metabolite_profiling_mass_spectrometry.txt	April 18 2023 17:39:07	
<input type="checkbox"/>	 i_Investigation.txt	April 18 2023 18:49:57	
<input type="checkbox"/>	 m_MTBLS3505_LC-MS_plasma_metabolite_profiling_v2_maf.tsv	April 18 2023 18:43:42	
<input type="checkbox"/>	 m_MTBLS3505_LC-MS_urine_metabolite_profiling_v2_maf.tsv	April 18 2023 18:43:42	
<input type="checkbox"/>	 s_MTBLS3505.txt	April 18 2023 16:53:14	

- S\_MTBLS\*\*\*\*.txt ファイル: 各サンプル・Rawデータに関する情報のテーブル
- i\_investigation.txt: Studyに関する各種メタデータ
- maf.tsvファイル：アノテーションされた代謝物のデータ行列

S\_MTBL5\*\*\*\*.txt

試料の種類・実験条件

サンプル

Sample name	Characteristic 1	Characteristic 2	Characteristic 3
C1	Mouse	Kidney	Control
C2	Mouse	Kidney	Control
N1	Mouse	Kidney	Drug Treated
N2	Mouse	Kidney	Drug Treated

maf.tsv

サンプル

代謝物

Database ID	inchi	Metabolite name	C1	C2	N1	N2
CHEBI:17263	InChI=...	Estrone	1.69E+08	7.0E+08	5.7E+08	7.2E+08
CHEBI:80658	InChI=...	Dihydrocortisol	1.52E+09	1.2E+09	3.6E+09	1.4E+09
CHEBI:732	InChI=...	Adipate semialdehyde	3.26E+09	1.9E+09	3.2E+09	1.4E+09
CHEBI:2490	InChI=...	3-Aminopentanedioate	1.69E+08	7.0E+08	3.2E+09	3.6E+06
CHEBI:28791	InChI=...	N-[(2S)-2-Amino-2-c...	1.69E+08	1.4E+09	7.1E+08	3.6E+09

- サンプル情報テーブル（S\_MTBL5\*\*\*.txt）と比較定量テーブル(maf.tsv)にメタボローム解析に必要な主要データが含まれている。
- 両ファイルを組み合わせることで容易に再解析が可能になる

# メタボロミクスデータ解析

- メタボロミクスでは膨大な代謝物データが得られるため、データを要約する解析が重要
- 多変量解析が良く用いられる
  - 主成分分析 (Principle component analysis)
  - 部分的最小二乗法 (Partial least squares)
- 生物学的な解釈のためのエンリッチメント解析

	試料A	試料B	試料C	試料D
metabolite A	50	100	90	80
Metabolite B	100	90	75	60
Metabolite C	30	60	100	90
Metabolite D	30	100	80	60

データ行列  
(比較定量テーブル)



## 多変量解析

- データの傾向・要約
- 重要な代謝物の選択



## エンリッチメント 解析

- 代謝パスウェイ特定
- 生物学的な解釈

# メタボロミクスデータ解析ツール



MSデータ処理

- MS-DIAL: 多様な装置の形式に対応したMSデータ解析ツール。ピーク検出・デコンボリューションやアノテーションで活躍。

MSデータ処理



- MZmine: マススペクトロメトリデータの解析のためのソフトウェア。ピーク検出、アライメント、データの視覚化、および多変量解析のサポートを含む。

解析・解釈



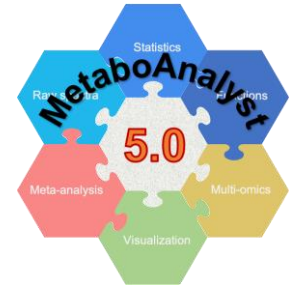
- Metabolomics workbench: レポジトリ・データベースに加えさまざまなメタボロミクス解析ツールを提供するウェブベースのプラットフォーム。多変量解析のサポートも含む。

解析・解釈



- MetaboAnalyst: ウェブベースのツールで、多変量解析・エンリッチメント解析を含む多数のデータ解析機能を提供する。APIやR言語での利用も可能

# MetaboAnalyst



- カナダのアルバータ大学のDavid Wishartと マギル大学のJianguo Xiaの研究グループにより運営されているフリーのWebサービス  
<http://www.metaboanalyst.ca/>
- Xia, J., Wishart, D.S. (2011), “Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst”, *Nature Protocols.*, 6, 743-760.
- Pang, Z., Chong, J., Zhou, G., Morais D., Chang, L., Barrette, M., Gauthier, C., Jacques, P.E., Li, S., and Xia, J. (2021) MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights *Nucl. Acids Res.* (doi: 10.1093/nar/gkab382)
- メタボロミクスデータの解析と解釈を支援するための包括的なウェブベースのツールであり、様々な統計的な解析手法を統合して、研究者がメタボロミクスデータを解釈するのを助ける。



# Metaboanalystデータ解析メニュー

Input Data Type	Available Modules (click on a module to proceed, or scroll down for more details)					
① Raw Spectra (mzML, mzXML or mzData)			LC-MS Spectra Processing			
② MS Peaks (peak list or intensity table)			Functional Analysis	Functional Meta-analysis		
③ Annotated Features (compound list or table)		Enrichment Analysis	Pathway Analysis	Joint-Pathway Analysis	Network Analysis	
④ Generic Format (.csv or .txt table files)	Statistical Analysis [one factor]	Statistical Analysis [metadata table]	Biomarker Analysis	Statistical Meta-analysis	Power Analysis	Other Utilities

多変量解析

<https://www.metaboanalyst.ca/>

- ① Raw spectra: ピーク処理前のRawデータ
- ② MS peaks: ピーク検出後の質量電荷比・シグナル強度情報のリスト
- ③ Annotated Features: アノテートされた化合物のリスト
- ④ Generic Format: csv, tsvフォーマットの様々なデータ

# データの準備（Metabolightsから）

S\_MTBLS\*\*\*\*.txt

サンプル

Sample name	Characteristic 1	Characteristic 2	Characteristic 3
C1	Mouse	Kidney	Control
C2	Mouse	Kidney	Control
N1	Mouse	Kidney	Drug Treated
N2	Mouse	Kidney	Drug Treated

群A

群B

サンプル名

maf.tsv

代謝物

Database ID	inchi	Metabolite name	C1	C2	N1	N2
CHEBI:17263	InChI=...	Estrone	1.69E+08	7.0E+08	5.7E+08	7.2E+08
CHEBI:80658	InChI=...	Dihydrocortisol	1.52E+09	1.2E+09	3.6E+09	1.4E+09
CHEBI:732	InChI=...	Adipate semialdehyde	3.26E+09	1.9E+09	3.2E+09	1.4E+09
CHEBI:2490	InChI=...	3-Aminopentanedioate	1.69E+08	7.0E+08	3.2E+09	3.6E+06
...	...	...	...	...	...	...

化合物名

多変量解析用  
テーブル

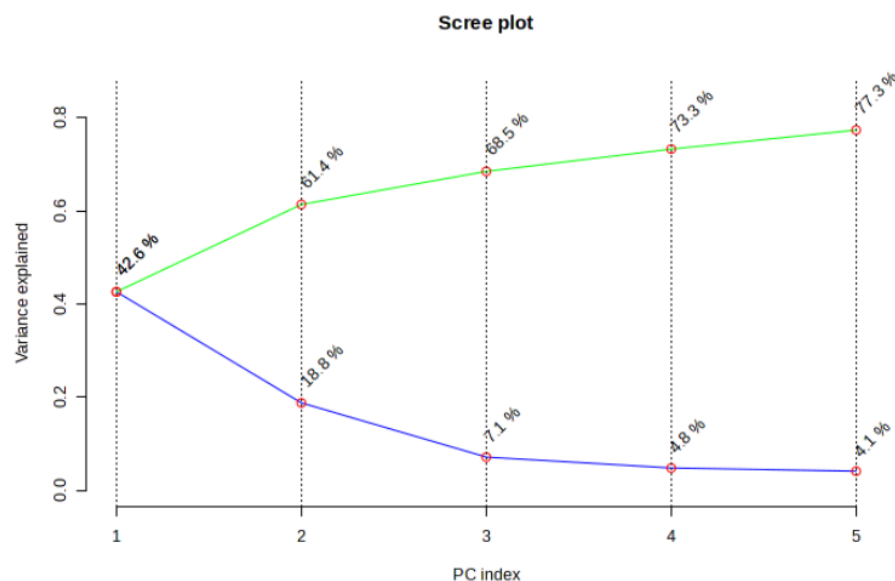
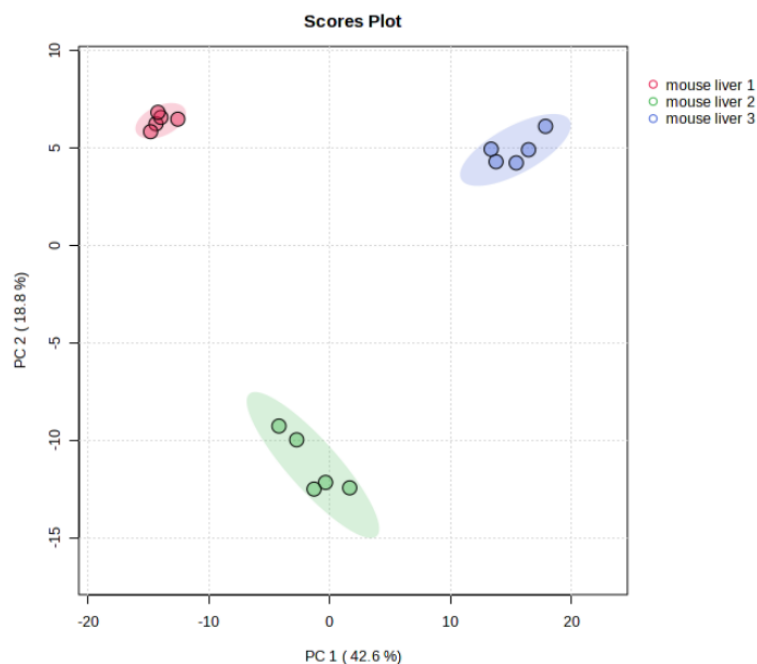
	C1	C2	N1	N2
Metabolite name	群A	群A	群B	群B
Estrone	1.69E+08	7.0E+08	5.7E+08	7.2E+08
Dihydrocortisol	1.52E+09	1.2E+09	3.6E+09	1.4E+09
Adipate semialdehyde	3.26E+09	1.9E+09	3.2E+09	1.4E+09
3-Aminopentanedioate	1.69E+08	7.0E+08	3.2E+09	3.6E+06
...	...	...	...	...

← サンプル名

← 群名

# メタボロミクスにおける代表的な多変量解析手法

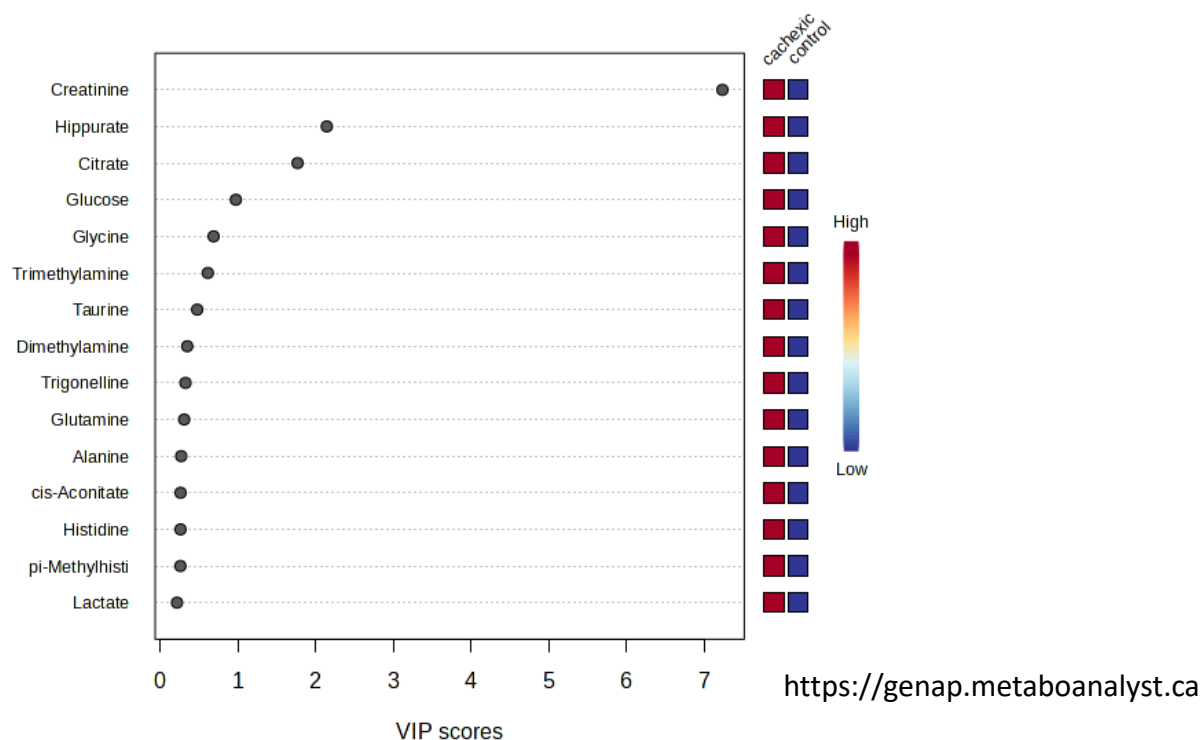
- 主成分分析 (Principle component analysis)
  - 大量の代謝物（変数）を説明する主成分を軸として次元を削減
  - サンプル群の代謝物プロファイルの傾向を可視化



<https://genap.metaboanalyst.ca>

# メタボロミクスにおける代表的な多変量解析手法

- 部分的最小二乗法 (Partial least squares)
  - 目的変数（サンプル情報等）と予測変数（代謝物群）の関係を潜在変数を介してモデリングを行う。
  - 目的変数への寄与が高い代謝物を選択



# 操作手順（多変量解析）

Please upload your data

A plain text file (.txt or .csv): ?

Data Type: ☐ Concentrations ☐ Spectral bins ☒ Peak intensities

Format:

Data File:

## ファイルアップロード

### Data Integrity Check:

1. Checking the class labels - at least three replicates are required in each class.
2. If the samples are paired, the pair labels must conform to the specified format.
3. The data (except class labels) must not contain non-numeric values.
4. The presence of missing values or features with constant values (i.e. all zeros).

### Data processing information:

Checking data content...passed.  
Samples are in columns and features in rows.  
The uploaded file is in comma separated values (.csv) format.  
The uploaded data file contains 12 (samples) by 400 (peaks(mz/rt)) data matrix.  
Samples are not paired.  
2 groups were detected in samples.  
Only English letters, numbers, underscore, hyphen and forward slash (/) are allowed.  
Other special characters or punctuations (if any) will be stripped off.  
All data values are numeric.  
A total of 0 (0%) missing values were detected.  
By default, missing values will be replaced by 1/5 of min positive values of their corresponding variables.  
Click the Proceed button if you accept the default practice.  
Or click the Missing Values button to use other methods.

## 入力データの確認

### Sample normalization

- ☒ None  
☐ Sample-specific normalization (i.e. weight, volume)   
☐ Normalization by sum  
☐ Normalization by median  
☐ Normalization by a reference sample (PQN)   
☐ Normalization by a pooled sample from group (group PQN)   
☐ Normalization by reference feature   
☐ Quantile normalization (suggested on microarray data)

### Data transformation

- ☒ None  
☐ Log transformation (base 10)  
☐ Square root transformation (square root of each value)  
☐ Cube root transformation (cube root of each value)

### Data scaling

- ☐ None  
☐ Mean centering (mean-centered only)  
☒ Auto scaling (mean-centered and divided by the standard deviation of each variable)  
☐ Pareto scaling (mean-centered and divided by the square root of the standard deviation of each variable)  
☐ Range scaling (mean-centered and divided by the range of each variable)

### Data scaling

- ☐ None  
☐ Mean centering (mean-centered only)  
☒ Auto scaling (mean-centered and divided by the standard deviation of each variable)  
☐ Pareto scaling (mean-centered and divided by the square root of the standard deviation of each variable)  
☐ Range scaling (mean-centered and divided by the range of each variable)

## Normalization/scaling

### Univariate Analysis

[Fold Change Analysis](#) [T-tests](#) [Volcano plot](#)

One-way Analysis of Variance (ANOVA)

[Correlation Heatmaps](#) [Pattern Search](#) [Correlation Networks \(DSPC\)](#)

### Advanced Significance Analysis

[Significance Analysis of Microarray \(and Metabolites\) \(SAM\)](#)

[Empirical Bayesian Analysis of Microarray \(and Metabolites\) \(EBAM\)](#)

### Chemometrics Analysis

[Principal Component Analysis \(PCA\)](#)

[Partial Least Squares - Discriminant Analysis \(PLS-DA\)](#)

[Sparse Partial Least Squares - Discriminant Analysis \(sPLS-DA\)](#)

[Orthogonal Partial Least Squares - Discriminant Analysis \(orthoPLS-DA\)](#)

### Cluster Analysis

Hierarchical Clustering: [Dendrogram](#) [Heatmaps](#)

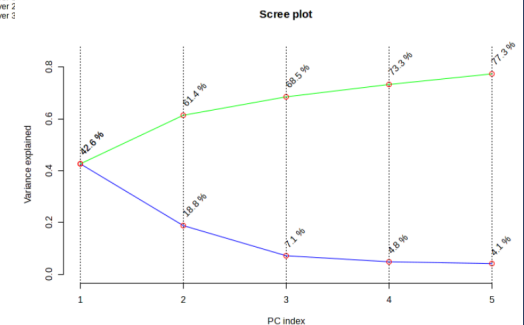
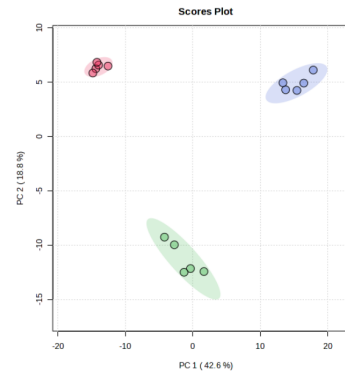
Partitional Clustering: [K-means](#) [Self Organizing Map \(SOM\)](#)

### Classification & Feature Selection

[Random Forest](#)

[Support Vector Machine \(SVM\)](#)

## PCA



## 解析結果のメニュー

# エンリッチメント解析

## Metabolite set enrichment analysis (MSEA)

- メタボロームデータと代謝パスウェイを関連づける手法
- t-検定などの統計的仮説検定は個々の代謝物について有意なものを選ぶが、エンリッチメント解析では代謝経路ごとに計算を行う。
- 「群間で変動している代謝物」と「代謝パスウェイ (metabolite set)」をマッチングし、特定の代謝パスウェイが統計的に有意に変動しているかを検証。

代謝経路X	変動した物質数	変動しない物質数	合計
代謝経路Xの物質数	10	5	15
代謝経路X以外の物質数	10	75	85
合計	20	80	100

- (統計的に厳密な解釈には注意が必要)

# MetaboAnalystでのエンリッチメント解析

## エンリッチメント解析

Input Data Type	Available Modules (click on a module to proceed, or scroll down for more details)					
Raw Spectra (mzML, mzXML or mzData)	LC-MS Spectra Processing					
MS Peaks (peak list or intensity table)			Functional Analysis	Functional Meta-analysis		
Annotated Features (compound list or table)		Enrichment Analysis	Pathway Analysis	Joint-Pathway Analysis	Network Analysis	
Generic Format (.csv or .txt table files)	Statistical Analysis [one factor]	Statistical Analysis [metadata table]	Biomarker Analysis	Statistical Meta-analysis	Power Analysis	Other Utilities

ID変換ツール

Please upload your data using one of the options below

Over Representation Analysis   Single Sample Profiling   Quantitative Enrichment Analysis

Please enter a one-column compound list:

Acetoacetic acid  
Beta-Alanine  
Creatine  
Dimethylglycine  
Fumaric acid  
Glycine  
Homocysteine  
L-Cysteine  
L-Isolucine  
L-Phenylalanine  
L-Serine  
L-Threonine  
L-Tyrosine  
L-Valine  
Phenylpyruvic acid  
Propionic acid  
Pyruvic acid

Input Type: Compound names

Feature Type: Metabolites

Try Example: ☐ None ☒ List 1 (metabolites) ☐ List 2 (lipids)

Submit

- (変動・有意差のあった) 代謝物リストをテキストボックスに入力
- 入力名はcompound name (HMDB)や各種ID が利用可能
- ID変換はID conversion toolが利用可能

# 入力情報の確認・標準化（Name/ID Standardization）

Query	Hit	HMDB	PubChem	KEGG	Details
Acetoacetic acid	Acetoacetic acid	<a href="#">HMDB0000060</a>	<a href="#">96</a>	<a href="#">C00164</a>	
Beta-Alanine	Beta-Alanine	<a href="#">HMDB0000056</a>	<a href="#">239</a>	<a href="#">C00099</a>	
Creatine	Creatine	<a href="#">HMDB0000064</a>	<a href="#">586</a>	<a href="#">C00300</a>	
Dimethylglycine	Dimethylglycine	<a href="#">HMDB0000092</a>	<a href="#">673</a>	<a href="#">C01026</a>	
Fumaric acid	Fumaric acid	<a href="#">HMDB0000134</a>	<a href="#">444972</a>	<a href="#">C00122</a>	
Glycine	Glycine	<a href="#">HMDB0000123</a>	<a href="#">750</a>	<a href="#">C00037</a>	
Homocysteine	Homocysteine	<a href="#">HMDB0000742</a>	<a href="#">778</a>	<a href="#">C00155</a>	
L-Cysteine	L-Cysteine	<a href="#">HMDB0000574</a>	<a href="#">5862</a>	<a href="#">C00097</a>	
<b>L-Isolucine</b>		-	-	-	<a href="#">View</a>
L-Phenylalanine	L-Phenylalanine	<a href="#">HMDB0000159</a>	<a href="#">6140</a>	<a href="#">C00079</a>	

<https://www.metaboanalyst.ca/>

- ユーザーの入力した代謝物情報の確認
- マッチしない（認識されない）代謝物は黄色でハイライトされるので view リンクから正しい名前を確認



# 解析で用いる代謝経路の分類指定 (Parameter Setting)

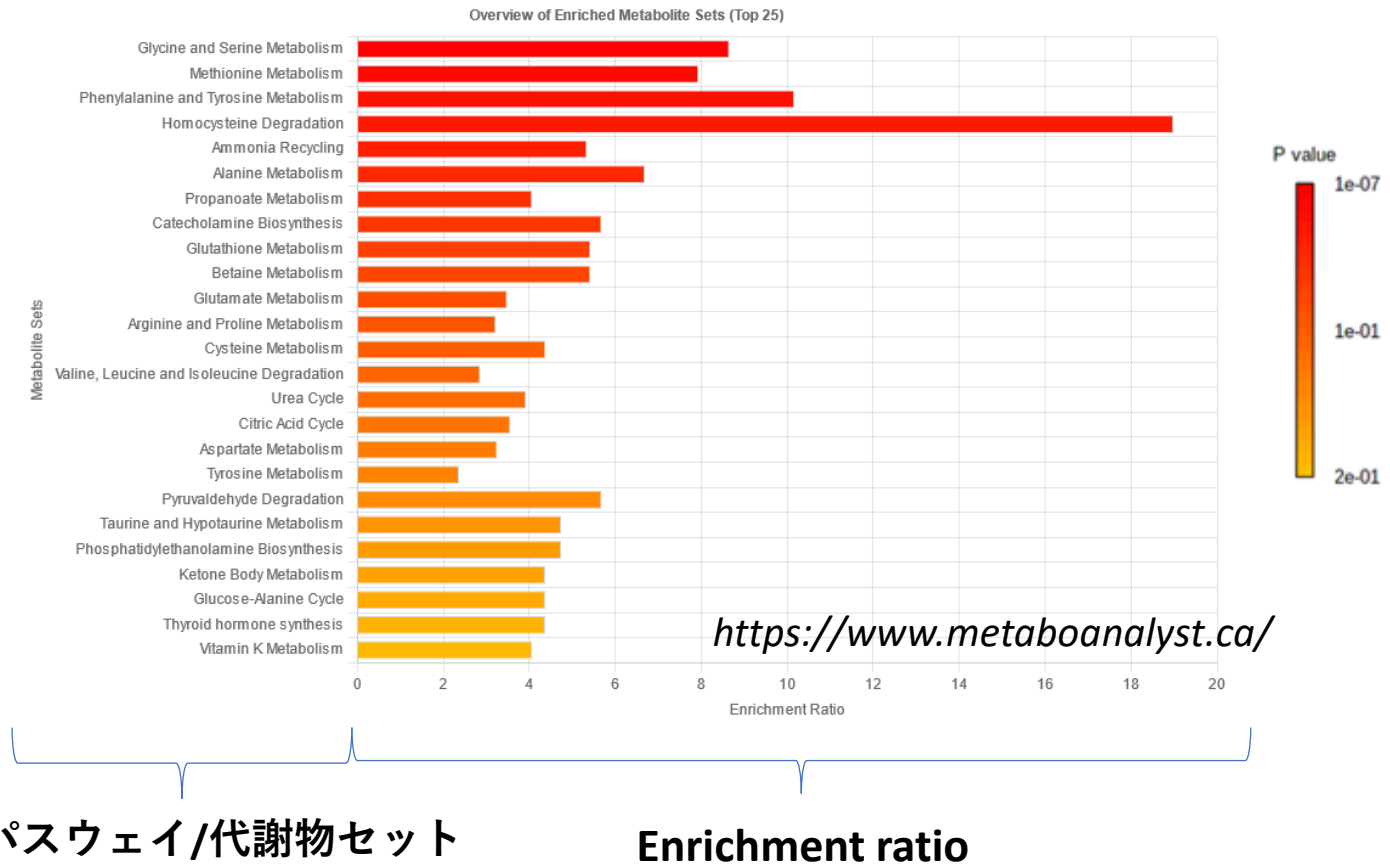
Please select a metabolite set library

Pathway based	<input checked="" type="radio"/> SMPDB <input type="radio"/> KEGG <input type="radio"/> Drug related	99 metabolite sets based on normal human metabolic pathways. 84 metabolite sets based on KEGG human metabolic pathways (Oct. 2019). 461 metabolite sets based on drug pathways from SMPDB.
Disease signatures	<input type="radio"/> Blood <input type="radio"/> Urine <input type="radio"/> CSF <input type="radio"/> Feces	344 metabolite sets reported in human blood. 384 metabolite sets reported in human urine. 166 metabolite sets reported in human cerebral spinal fluid (CSF). 44 metabolite sets reported in human feces.
Chemical structures	<input type="radio"/> Super-class <input type="radio"/> Main-class <input type="radio"/> Sub-class	35 super chemical class metabolite sets or lipid sets 464 main chemical class metabolite sets or lipid sets 1072 sub chemical class metabolite sets or lipid sets
Other types	<input type="radio"/> SNPs <input type="radio"/> Predicted <input type="radio"/> Locations	4,598 metabolite sets based on their associations with SNPs loci. 912 metabolic sets predicted to change in the case of dysfunctional enzymes. 73 metabolite sets based on organ, tissue, and subcellular localizations.
Self defined	<input type="radio"/> Upload here	define your own customized metabolite sets

<https://www.metaboanalyst.ca/>

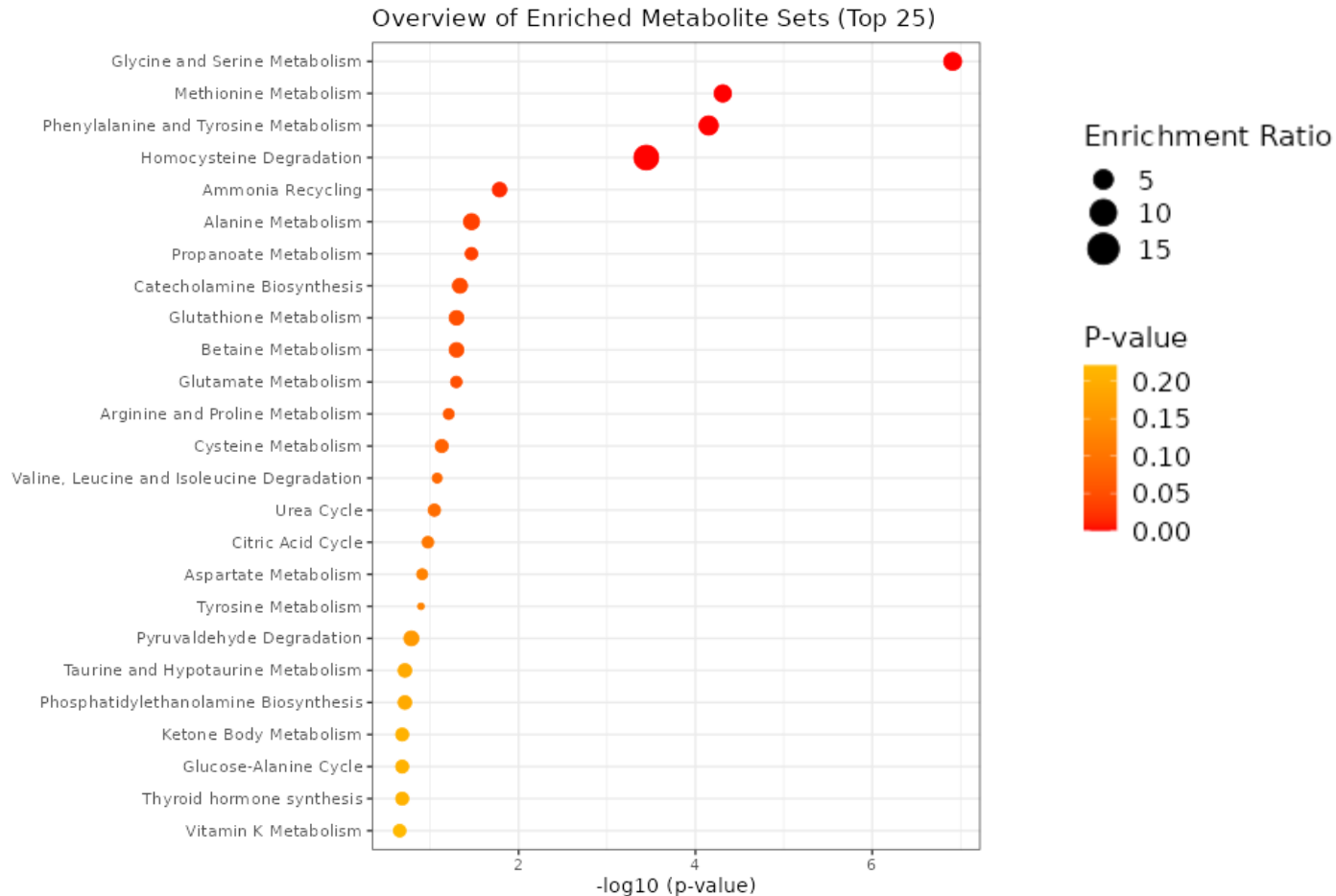
- 目的に合った代謝経路の分類を選択する。
- 一般的なメタボロミクスの場合はPathway-associated metabolite sets (SMPDB)を選択
- 他のオミクス解析でKEGG、SNPs、組織等の情報と連携させる場合は該当の物を選択
- 自分で作成した代謝経路の分類・代謝物のセットを利用することも可能

# エンリッチメント解析結果①



- **Enrichment ratio:** 特定の代謝パスウェイ(代謝物セット)内の代謝物の頻度と、全体のデータセット内のその頻度との比率。

## エンリッチメント解析結果②



<https://www.metaboanalyst.ca/>

エンリッチメント解析結果のP-valueとEnrichment Ratioを可視化

# 解析結果の詳細

Metabolite Set	Total	Hits	Expect	P value	Holm P	FDR	Details
Glycine and Serine Metabolism	59	9	1.04	1.23E-7	1.21E-5	1.21E-5	<a href="#">View</a>
Methionine Metabolism	43	6	0.756	4.9E-5	0.00476	0.00231	<a href="#">View</a>
Phenylalanine and Tyrosine Metabolism	28	5	0.492	7.08E-5	0.00679	0.00231	<a href="#">View</a>
Homocysteine Degradation	9	3	0.158	3.59E-4	0.0341	0.00881	<a href="#">View</a>
Ammonia Recycling	32	3	0.562	0.0164	1.0	0.322	<a href="#">View</a>
Alanine Metabolism	17	2	0.299	0.034	1.0	0.451	<a href="#">View</a>
Propanoate Metabolism	42	3	0.738	0.0341	1.0	0.451	<a href="#">View</a>

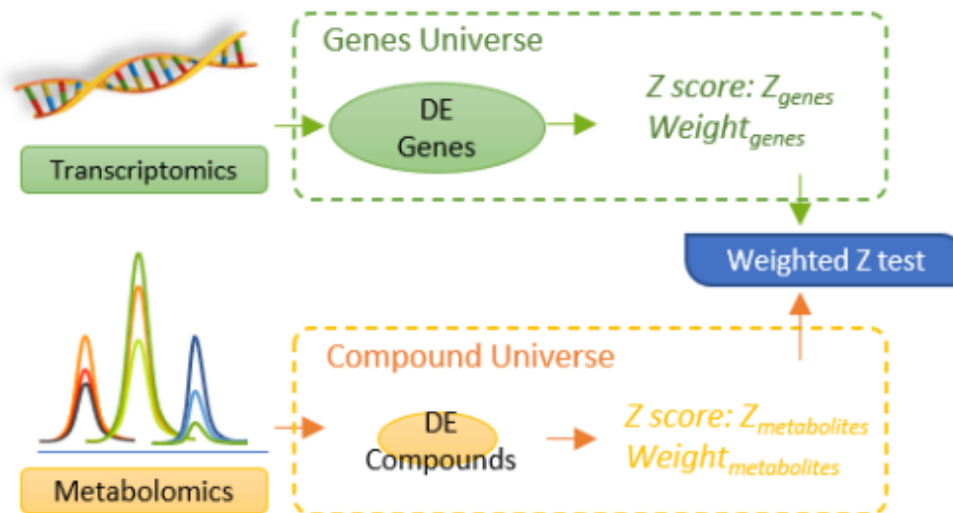
<https://www.metaboanalyst.ca/>

- Holm P : Holm法による多重比較補正を行ったp-value
- FDR: 想定される偽陽性の割合
- Viewリンクから代謝パスウェイ/代謝物セットの詳細が確認できる

Set Name	Metabolites	References
Glycine and Serine Metabolism	2-Ketobutyric acid; Betaine; Adenosine monophosphate; Ammonia; <b>Creatine</b> ; <b>Dimethylglycine</b> ; L-Cystathionine; Glyoxylic acid; <b>Glycine</b> ; Guanidoacetic acid; Glyceric acid; L-Glutamic acid; L-Alanine; <b>L-Threonine</b> ; <b>L-Serine</b> ; Oxoglutaric acid; Ornithine; Orotidylic acid; <b>Pyruvic acid</b> ; Pyrophosphate; <b>Sarcosine</b> ; Phosphoserine; L-Arginine; Adenosine triphosphate; Magnesium; <b>L-Cysteine</b> ; L-Methionine; <b>Homocysteine</b> ; 3-Phosphoglyceric acid; NAD; S-Adenosylhomocysteine; Succinyl-CoA; Phosphohydroxypyruvic acid; 5-Aminolevulinic acid; Pyruvaldehyde; S-Adenosylmethionine; Acetyl-CoA; FAD; Zinc (II) ion; ADP; Hydroxypyruvic acid; Oxygen; Coenzyme A; Formaldehyde; Phosphate; (R)-lipoic acid; NADH; Pyridoxal 5'-phosphate; 5,10-Methylene-THF; Tetrahydrofolic acid; Carbon dioxide; Water; Aminoacetone; Hydrogen peroxide; D-Serine; L-2-Amino-3-oxobutanoic acid; Dihydrolipoate; 8-((Aminomethyl)sulfonyl)-6-sulfonyloctanoic acid; (S)-lipoic acid	<a href="#">SMPDB</a> <a href="#">KEGG</a>

# マルチオミクス（マルチセット）のデータ解析

- マルチセットPLS
  - 教師有マルチセットPLS-ROG  
(Yamamoto, H., *Journal of Chemometrics*, 31(3) (2017) e2883)
- Joint Pathway Analysis (MetaboAnalyst)
  - マルチセット多変量解析の結果で統計的に優位な遺伝子・タンパク質・代謝物に対してエンリッチメント解析を行う。



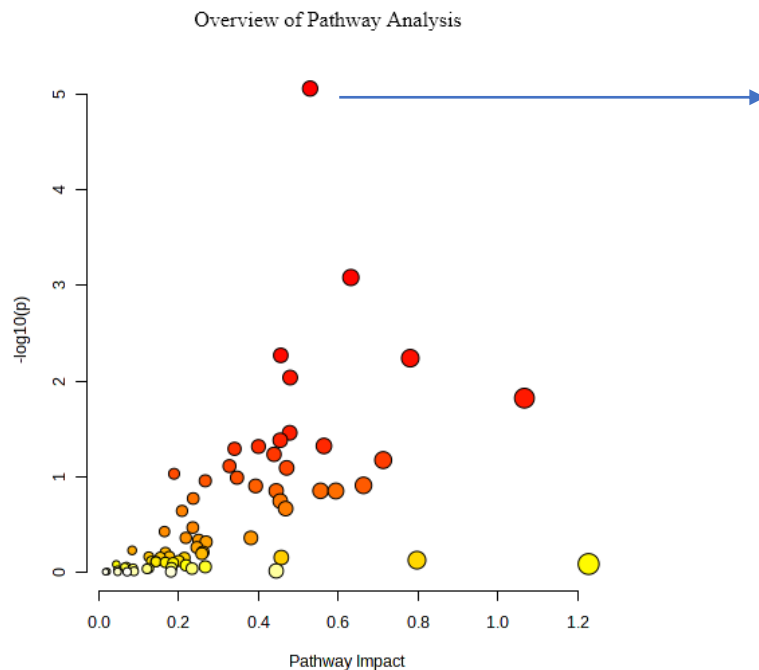
from MetaboAnalyst tutorial page  
(<https://www.metaboanalyst.ca/MetaboAnalyst/docs/Tutorials.xhtml>)

# MetaboanalystにおけるJoint Pathway Analysis

- **Menu**

Input Data Type	Available Modules (click on a module to proceed, or scroll down for more details)					
Raw Spectra (mzML, mzXML or mzData)				LC-MS Spectra Processing		
MS Peaks (peak list or intensity table)			Functional Analysis	Functional Meta-analysis		
Annotated Features (compound list or table)		Enrichment Analysis	Pathway Analysis	Joint-Pathway Analysis	Network Analysis	
Generic Format (csv or txt batch files)	Statistical Analysis (one factor)	Statistical Analysis (metadata table)	Biomarker Analysis	Statistical Meta-analysis	Power Analysis	Other Utilities

## ● 結果



- データ入力

Organism:

----- Not Specified -----

Metabolomics Type:

Targeted (compound list) ▾

Genes/proteins with optional fold changes

Genes/proteins

Compound list with optional fold changes

Metabolites

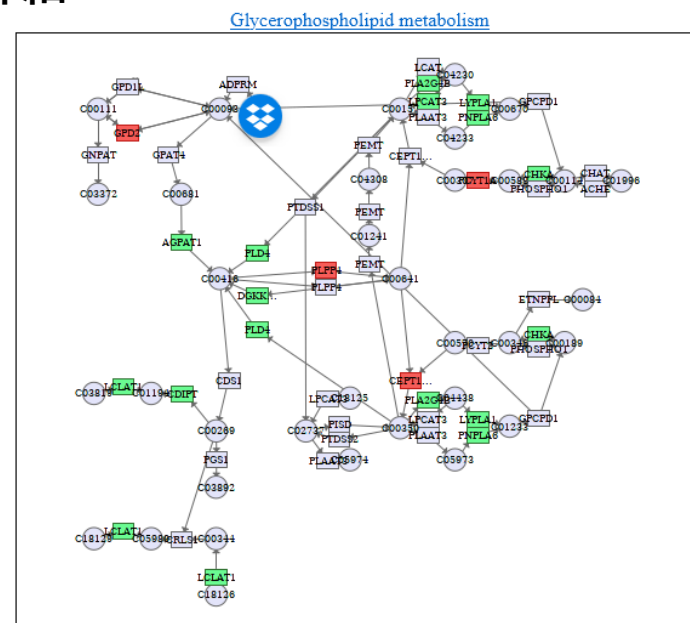
ID Type:

--- Not Specified --- ▾

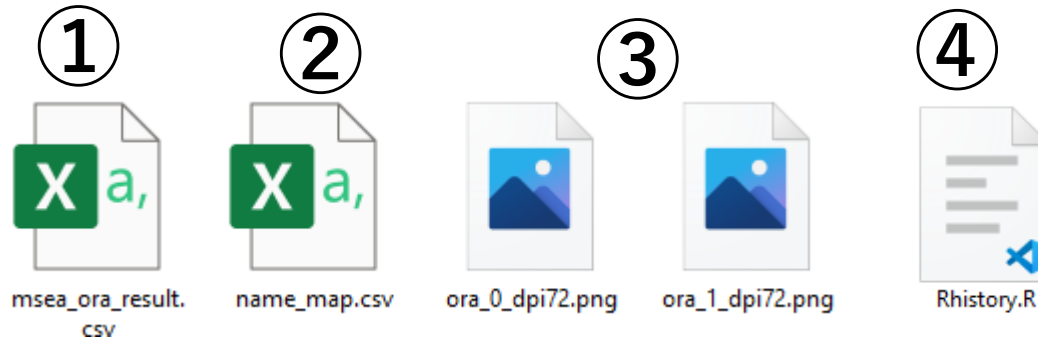
ID Type:

--- Not Specified --- ▾

- 詳細



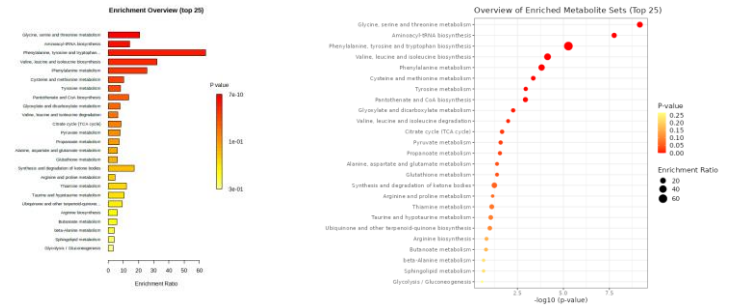
# 解析結果のダウンロード



## ①エンリッチメント解析結果

	A	B	C	D	E	F	G
1		total	expected	hits	Raw p	Holm p	FDR
2	Glycine and Serine Metabo	59	1.04	9	1.23E-07	1.21E-05	1.21E-05
3	Methionine Metabolism	43	0.756	6	4.90E-05	0.00476	0.00231
4	Phenylalanine and Tyrosine	28	0.492	5	7.08E-05	0.00679	0.00231
5	Homocysteine Degradation	9	0.158	3	0.00036	0.0341	0.00881
6	Ammonia Recycling	32	0.562	3	0.0164	1	0.322
7	Alanine Metabolism	17	0.299	2	0.034	1	0.451
8	Propanoate Metabolism	42	0.738	3	0.0341	1	0.451
9	Catecholamine Biosynthesis	20	0.352	2	0.046	1	0.451
10	Glutathione Metabolism	21	0.369	2	0.0503	1	0.451
11	Betaine Metabolism	21	0.369	2	0.0503	1	0.451

## ③各種画像ファイル



## ②使用した代謝物リスト・マッチ結果

	A	B	C	D	E	F	G
1	Query	Match	HMDB	PubChem	KEGG	SMILES	Commer
2	Acetoacetic acid	Acetoacetic acid	HMDB000006	96	C00164	CC(=O)CC	1
3	Beta-Alanine	Beta-Alanine	HMDB000005	239	C00099	C(CN)C(=O)	1
4	Creatine	Creatine	HMDB000006	586	C00300	CN(CC(=O)	1
5	Dimethylglycine	Dimethylglycine	HMDB000005	673	C01026	CN(C)CC(=O)	1
6	Fumaric acid	Fumaric acid	HMDB000013	444972	C00122	C(=C/C(=O)	1
7	Glycine	Glycine	HMDB000012	750	C00037	C(C(=O)O)	1
8	Homocysteine	Homocysteine	HMDB000074	778	C00155	C(CS)C(C(=O)	1
9	L-Cysteine	L-Cysteine	HMDB000057	5862	C00097	C([C@@H](N)	1

## ④Rコード

```

1 # PID of current job: 2344740
2 mSet<-InitDataObjects("conc", "msetora", FALSE)
3 compd.vec<-c("Acetoacetic acid", "Beta-Alanine", "Creatine", "Dimethylglyci
4 mSet<-Setup.MapData(mSet, compd.vec);
5 mSet<-CrossReferencing(mSet, "name");
6 mSet<-CreateMappingResultTable(mSet)
7 mSet<-PerformDetailMatch(mSet, "L-Isolucine");
8 mSet<-GetCandidateList(mSet);
9 mSet<-SetCandidate(mSet, "L-Isolucine", "L-Isoleucine");
10 mSet<-SetMetabolomeFilter(mSet, F);
11 mSet<-SetCurrentMsetLib(mSet, "kegg_pathway", 2);
12 mSet<-CalculateHyperScore(mSet)
13 mSet<-PlotORA(mSet, "ora_0_", "net", "png", 72, width=NA)
14 mSet<-PlotEnrichDotPlot(mSet, "ora", "ora_dot_0_", "png", 72, width=NA)
15 mSet<-CalculateHyperScore(mSet)
16 mSet<-PlotORA(mSet, "ora_1_", "net", "png", 72, width=NA)
17 mSet<-PlotEnrichDotPlot(mSet, "ora", "ora_dot_1_", "png", 72, width=NA)
18 mSet<-SetMetabolomeFilter(mSet, F);

```

## Summary

- メタボロミクスデータ解析の流れ
  - Rawデータ→MSデータ処理 → データ行列 →解析
- 公共データレポジトリ
  - Rawデータに加え各種ファイルを取得可能
  - 構造化されたデータファイルから効率的に再解析を行う
- データ解析ツール
  - 多変量解析でデータの傾向と重要な代謝物を知る
  - エンリッチメント解析で代謝パスウェイと生物学的な解釈へ



# 質量分析インフォマティクス

- 質量分析インフォマティクス研究会 (<http://ms-bio.info/>)
  - 「質量分析法を用いた研究分野（特にオミクス研究）のデータ解析」の研究者の交流の場
  - ハッカソン（国内版バイオハッカソンと共同開催）・ソフトウェア講習会・ミートアップ（テーマごとのディスカッション・情報交換）
- メタボロミクス実践ガイド（書籍）
  - 試料調整、機器分析、データ解析まで網羅  
<https://www.yodosha.co.jp/yodobook/book/9784758122511/>
- 質量分析インフォマティクス入門
  - 質量分析インフォマティクスに関する各種情報を紹介  
<https://mass-spec-info.github.io/metaboguide/>  
（google検索：“質量分析インフォマティクス” & “github”）





# 主成分負荷量とは

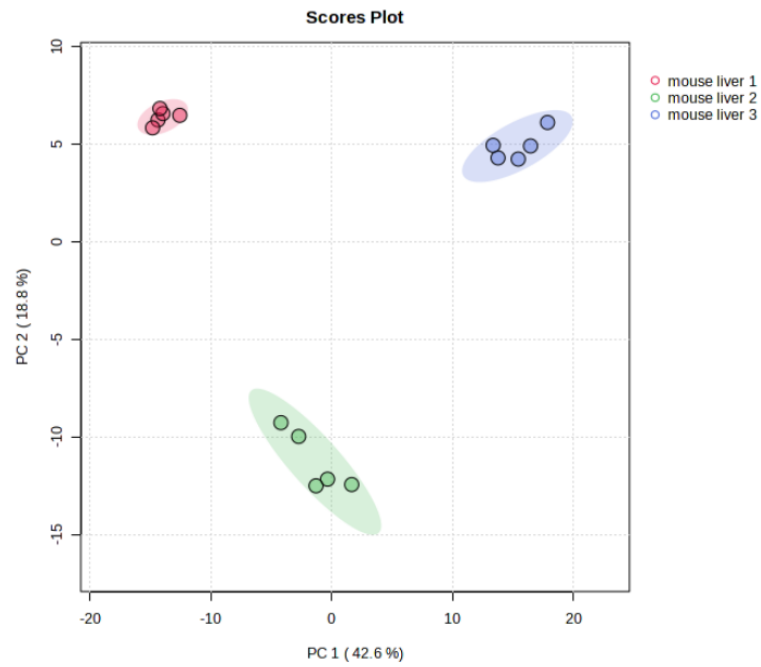
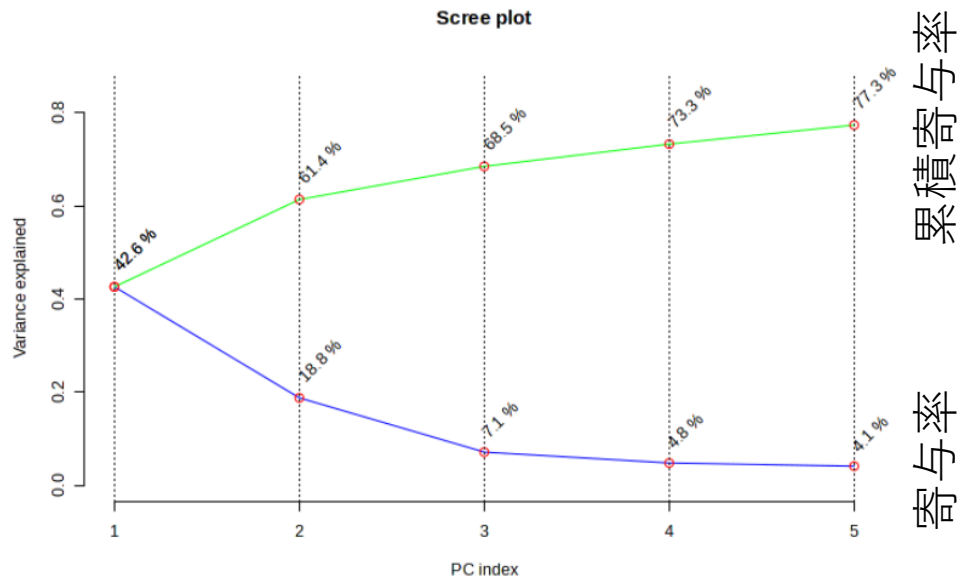
## • 主成分負荷量とは？

- MetaboAnalystの主成分分析で得られるloadingは、**主成分負荷量**ではなく、**主成分係数(固有ベクトル・重み係数)**
  - loadingと言っても、ソフトウェアによって出力される値が異なる  
SPSSは**主成分負荷量**、SIMCAはMetaboAnalystと同じく**主成分係数**がloadingという名前になっている
- **主成分係数**ではなく、**主成分負荷量**を利用するメリット
  - **主成分負荷量**は、主成分スコアと各代謝物の**相関係数**で定義され、その値の大きさを利用して重要な代謝物を選ぶことが出来る。
    - 例えば、相関係数が0.7以上だと相関が強い、など。
  - **主成分負荷量**は仮説検定を行うことが出来、 $p$ -valueが計算可能
    - Yamamoto et al., BMC Bioinformatics, (2014) 15(1):51. を参照

## • 主成分負荷量の計算方法

- **主成分係数**から**主成分負荷量**に変換する方法を紹介

変動（分散）を  
どれだけ説明しているか



MetaboAnalystの主成分分析で得られるloadingは、**主成分負荷量(各変数(特徴量)が主成分にどの程度寄与しているかを示す数値)**ではなく、**主成分係数(固有ベクトル・重み係数)**

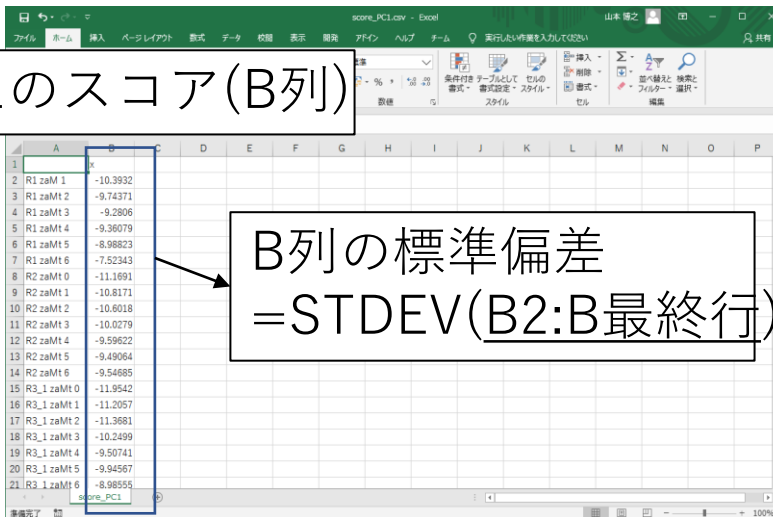
## 主成分負荷量

$$= \text{主成分スコアの標準偏差} \times \text{主成分係数} \\ (\text{分散の平方根})$$

(ただしAutoscalingされている場合のみ)

主成分スコア、**pca\_score.csv**  
(MetaboAnalystからダウンロードしたもの)

PC1のスコア(B列)



B列の標準偏差  
=STDEV(B2:B最終行)

	A	B
1		
2	R1 zaM 1	-10.3932
3	R1 zaM 2	-9.74371
4	R1 zaM 3	-9.2806
5	R1 zaM 4	-9.36079
6	R1 zaM 5	-8.98823
7	R1 zaM 6	-7.52343
8	R2 zaM 0	-11.1891
9	R2 zaM 1	-10.8171
10	R2 zaM 2	-10.6018
11	R2 zaM 3	-10.0279
12	R2 zaM 4	-9.59622
13	R2 zaM 5	-9.49064
14	R2 zaM 6	-9.54885
15	R3_1 zaM 0	-11.9542
16	R3_1 zaM 1	-11.2057
17	R3_1 zaM 2	-11.3681
18	R3_1 zaM 3	-10.2499
19	R3_1 zaM 4	-9.50741
20	R3_1 zaM 5	-9.94567
21	R3_1 zaM 6	-8.98555

**主成分係数**  
(固有ベクトル・重み係数)

**pca\_loadings.csv**

主成分係数に  
同じ値を掛ける

参考：<https://github.com/hiroyuki Yamamoto/mseapca/blob/master/protocols/主成分負荷量.xlsx>

## PLS output

- **VIP (Variable Importance in the Projection)** : 変数（代謝物のシグナル強度）の重要性を示す指標として使用される。この値が大きいほど、その変数はモデルの分類能力に大きく寄与していることを示す。一般的には、VIPの値が1より大きい場合、その変数はモデルにおいて重要であると考えられる。

- **R2**: 高ければ高いほど良いが、**Q2**との関係も考慮する必要がある。

- **Q2**: 一般に、0.5以上であれば予測の品質は良好とされている。

しかし、以下の点も注意が必要：

- **R2**が非常に高く、**Q2**が低い場合、モデルはオーバーフィッティングしている可能性が高い。

- 統計的検証（例えば、パーミュテーションテスト）を行い、モデルのロバスト性を確認することが推奨。

- **R2**や**Q2**の値だけに依存せず、生物学的な解釈や他の統計的評価と組み合わせてモデルを評価することが重要。

代謝経路X	変動した物質数	変動しない物質数	合計
代謝経路Xの物質数	10	5	15
代謝経路X以外の物質数	10	75	85
合計	20	80	100

```
from scipy.stats import fisher_exact
```

```
# 2x2の分割表
```

```
table = [[10, 5], [10, 75]]
```

```
# フィッシャーの正確確率検定
```

```
odds_ratio, p_value = fisher_exact(table)
```

```
print(p_value)
```

```
1.8622426836555598e-05
```

Pathway impact here represents a combination of the centrality and pathway enrichment results; higher impact values represent the relative importance of the pathway; the size of the circle indicates the impact of the pathway while the color represents the significance (the more intense the red color, the lower the p).



3群のPLS-DAの手順：

1.ダミー変数の作成：3群のクラスを表現するために、ダミー変数（またはインジケータ変数）を使ってエンコードします。3群が A, B, C であるとしたら、以下のように変換します：  
makefile

1.A: [1, 0, 0] B: [0, 1, 0] C: [0, 0, 1]

このように、各サンプルは3つの変数のいずれかを持つことになります。これにより、3群の情報を目的変数の行列として表現できます。

**2.PLS-DAの実行:** このエンコーディングを使用して、部分最小二乗法を実行します。ここで、X変数（予測変数）とY変数（エンコードされたクラス情報）の共変性を最大化する成分を見つけ出すことが目的となります。

**3.クラス予測:** 新しいサンプルのクラスを予測するためには、PLSモデルを使用して予測されるY変数の値（3つのダミー変数の値）を取得し、最も高い値を持つ変数のクラスを選択します。

**4.モデルの評価:** 交差検証を使用してモデルの予測精度を評価します。オーバーフィッティングのリスクを軽減するため、交差検証は非常に重要です。

**5.変数の重要性:** PLS-DAでは、Variable Importance in the Projection (VIP) スコアを用いて、各変数のクラス分別における重要性を評価できます。

要するに、3群以上のPLS-DAでは、群の情報をダミー変数としてエンコードすることで、複数のクラスを目的変数として扱うことができます。このアプローチにより、PLSのフレームワーク内で多クラスの分別問題を解決することができます。