Single sample enrichment analysis for mass spectrometry-based omics data

Hiroyuki Yamamoto

h.yama2396@gmail.com

Japan Computational Mass Spectrometry (JCompMS) group

Abstract

In omics studies using mass spectrometry, such as metabolomics and proteomics, it is challenging to detect all metabolites or proteins. This difficulty can potentially introduce bias into traditional enrichment analyses. To address this issue, this study proposes an analytical method using single-sample enrichment analysis. Unlike conventional methods, our approach leverages information from both detected compounds and those outside the detection scope. It is particularly useful for omics data from mass spectrometry with many missing values and where comprehensively covering all compounds is challenging.

Keyword single sample enrichment analysis, missing value, metabolomics

質量分析オミックスデータのための
シングルサンプルエンリッチメント解析

質量分析インフォマティクス研究会　山本　博之　h.yama2396@gmail.com

Abstract メタボロミクスやプロテオミクスをはじめとする質量分析を用いたオミックス研究では、全ての代謝物もしくはタンパク質を検出することが難しく、従来のエンリッチメント解析を行う際に、その結果に偏りが生じる可能性がある。この問題を解決するために、本研究ではシングルサンプルエンリッチメント解析を用いた解析方法を提案する。本手法は、検出された化合物とそれ以外の化合物の情報を利用している点が従来とは異なり、全ての化合物を完全に網羅することが難しく欠損値が多い質量分析を用いたオミックスデータに対して特に有用である。

1. Introduction

メタボロミクスやプロテオミクスをはじめとするオミックス研究では、多変量解析や統計的仮説検定などを用いて重要な代謝物を選び出し、代謝パスウェイをはじめとする生物学的に意味のある分類との関連付けを行うことにより結果の解釈が行われている。エンリッチメント解析の手法として Over representation analysis (ORA)[1]や Gene set enrichment analysis (GSEA)[2]をはじめとして様々な手法が提案されている。

GSVA (Gene Set Variation Analysis)[3]をはじめとする Single sample enrichment analysis (SSEA)[4]では、それぞれのサンプルに対して遺伝子発現量の大きさで並び替えた後、その値の大きさに対してエンリッチメント解析を行う。得られたエンリッチメントスコアを行列データとし、その後、統計的仮説検定や多変量解析、機械学習などの各種解析を行う。一方メタボロミクスでは、面積値もしくは相対面積値がデータとして利用されることが多く、その値は代謝物の濃度と質量分析の感度の両方に依存し、各物質の濃度の大小関係を直接表しているとは言えないことから、その値の大きさで並べることは必ずしも適切とは言えない。

さらに質量分析を用いたオミックス研究では、遺伝子発現とは異なり全ての物質を網羅的に解析することは出来ない。その結果、検出された代謝物が多く含まれる代謝物セット(代謝パスウェイなど)と、少数しか含まれない代謝物セットが混在し、実際にエンリッチメント解析を行った際に、各代謝物セットに対する検出物質数のバイアスを考慮して結果を解釈する必要があるといった課題が残されていた。

そこで本研究では、新しい SSEA として、サンプル毎に検出された物質とそれ以外を 2 値として ORA を行った。これにより、検出された物質が多い代謝物セットを大きく、検出された物質が少ない代謝物セットを小さく表すスコアをデータとし、代謝物セットの偏りをスコアに組み込むことで、各代謝物セットに対する検出物質数の偏りの問題を回避出来る可能性がある。実際にメタボロームデータに本手法を適用し、従来法[5]と比較した結果に

ついて紹介する。

2. Methods

代謝パスウェイなどの代謝物セットに対して、サンプル毎にエンリッチメント解析を行う。具体的には、代謝物が検出されたかどうかと、代謝物セットに含まれるかどうか、の2つの項目に対して該当する代謝物の数をカウントし、2×2のクロス集計表を作成し、フィッシャーの正確確率検定を行う。例えば、解糖系の代謝物群が多く検出され、かつ解糖系以外の代謝物群について検出された代謝物が少ない場合は、p値が小さくなる。本研究では、p値の対数を取り、マイナスの符号を付けたものをSSEAスコアとする。これにより、特定の代謝物セットに含まれている代謝物が多く検出されていればスコアの値は高くなり、逆に特定の代謝物セットに含まれる代謝物のうち検出されたものが少なければ、スコアの値が小さくなる統計量となる。このスコアをサンプル毎に計算し、この得られたスコアを統計的仮説検定、多変量解析、機械学習などの各種解析に利用することが出来る。また、タンパク質と代謝物のそれぞれのSSEAスコアの相関係数のHeatmapから関連の高い組み合わせを見つけることで、マルチオミックス解析に利用することも出来る。

本研究で提案したORAによるSSEAは、検出された代謝物と検出されなかった代謝物を選定する基準に影響を受ける。本研究で用いたデータは信号強度がS/N=3未満の場合に検出されなかったと定義しているが、S/Nの情報を利用できる場合には、閾値としてS/Nを他の値、例えば10などを設定することも出来る。またS/Nが利用できない場合でも、z変換後のz-scoreを用いて代謝物毎に信号強度の閾値を設定し、閾値以上の値を示す代謝物を検出出来た代謝物とみなして解析を行うことも出来る。

本研究では、絶食マウスのメタボロームデータ[5]、COVID-19のプロテオームデータ[6]を用いて解析を行った。データはloadingsパッケージのデータ、メタボロームデータ用のパスウェイは独自に作成し、プロテオームデータ用のパスウェイはPathBankデータベースを採用し実際のパスウェイを取得するためRのBioconductorパッケージAHPathbankDbs (https://bioconductor.org/packages/release/data/annotation/html/AHPathbankDbs.html)を用いた。またCOVID-19のメタボロームデータとプロテオームデータのマルチオミックス解析では、タンパク質のアノテーションはUniProt IDが付与されているのでそれをそのまま用いた。一方で代謝物のアノテーションは、物質名のみ付与されていたことから、MetaboAnalystのCompound ID Conversionを用いてHMDB IDに変換したもののみ用いた。パスウェイデータベースは、これまでと同様にPathBankを用いた。プロテオーム、メタボロームいずれも個々のタンパク質、代謝物の生合成に関するパスウェイ(Protein Synthesis: AlanineやPhosphatidylcholine Biosynthesis PC(14:0/14:0)など)を除外したものを用いた。

以上の全ての計算はRを用いて行い、mseapcaパッケージとして公開している(https://cran.r-project.org/web/packages/mseapca)。また主成分負荷量の計算にはloadingsパッケ

ージ(https://cran.r-project.org/web/packages/loadings/)を用いた。

3. Results and Discussion

3-1. 従来の ORA とその問題点

ORA による従来のエンリッチメント解析は、多変量解析や統計的仮説検定などから得られた統計的に有意な代謝物に対して行われる。ORA は遺伝子発現データで Gene Ontology 解析として広く用いられているが、遺伝子発現データと質量分析を用いたオミックス研究であるメタボロミクスやプロテオミクスとの最大の違いは、検出できない代謝物もしくはタンパク質が数多く存在することである。これは、ORA を用いたエンリッチメント解析をメタボロミクスもしくはプロテオミクスに適用する際に大きく影響を及ぼす。

質量分析を用いたオミックス研究において ORA を用いたエンリッチメント解析の問題点を示すために、絶食マウスの肝臓のメタボロームデータを用いて、エンリッチメント解析の 2 つの方法の比較結果を紹介する。1 つは、遺伝子発現データと同様にパスウェイ情報をそのまま用いて ORA によるエンリッチメント解析を行った場合、もう一つはエンリッチメント解析のパスウェイに含まれる代謝物の中で、検出された代謝物のみを用いて ORA によるエンリッチメント解析を行った場合の結果である。Table1 にそれぞれの方法で解析した結果を示す。

Table1 ORA によるエンリッチメント解析の結果

Metabolic pathways	all			detected		
	p-value	q-value		p-value	q-value	
Glycolysis	0.0000	0.0000	*	0.0000	0.0006	*
Pentose phosphate pathway	0.0034	0.0365	*	0.0340	0.2211	
TCA cycle	0.0042	0.0365	*	0.0046	0.0593	
Glutamic acid and glutamine metabolism	0.4552	1.0000		0.5046	0.8746	
Alanine, aspartic acid and asparagine metabolism	0.0499	0.3246		0.3047	0.7922	
Lysine metabolism	0.9894	1.0000		0.8347	1.0000	
Valine, leucine and isoleucine metabolism	0.9491	1.0000		0.7828	1.0000	
Glycine, serine and threonine metabolism	0.2727	0.7411		0.2650	0.7922	
Cysteine metabolism	0.2134	0.6936		0.0929	0.3450	
Methionine metabolism	0.9677	1.0000		0.8523	1.0000	
Shikimic acid metabolism	0.8918	1.0000		0.5046	0.8746	
Histidine metabolism	1.0000	1.0000		1.0000	1.0000	
Urea cycle	0.0791	0.4114		0.0565	0.2937	
Proline metabolism	0.5631	1.0000		0.2836	0.7922	

Polyamine metabolism	0.1029	0.4458		0.0340	0.2211	
Tryptophan metabolism	1.0000	1.0000		0.9321	1.0000	
Tyrosine metabolism	0.9633	1.0000		0.0813	0.3450	
beta-alanine metabolism	0.8023	1.0000		0.8738	1.0000	
Taurine metabolism	0.2850	0.7411		0.3916	0.8746	
Creatine metabolism	0.1794	0.6664		0.4886	0.8746	
Purine metabolism	0.9999	1.0000		0.9980	1.0000	
Pyrimidine metabolism	0.9827	1.0000		0.9274	1.0000	
Ribonucleotide metabolism	0.9311	1.0000		0.9906	1.0000	
Deoxyribonucleotide	1.0000	1.0000		1.0000	1.0000	
Conjugated bile acid	1.0000	1.0000		1.0000	1.0000	
Nicotinic acid metaboilsm	0.6822	1.0000		0.4770	0.8746	

この結果を見ると、前者は $q<0.05$ で統計的に有意な代謝パスウェイは Glycolysis、Pentose phosphate pathway、TCA cycle の 3 つであるのに対し、後者は $q<0.05$ で統計的に有意な代謝パスウェイは Glycolysis のみであった。この結果より、前者の方が統計的に有意になりやすく、後者の方が統計的に有意になりにくい傾向にあることが分かる。例えば Glycolysis に着目した場合、前者と後者のクロス集計表はそれぞれ

Table2　Glycolysis の 2×2 クロス集計表

all	有意	有意でない	合計
Glycolysis	9	1	10
その他	80	563	643
合計	89	564	653

detected	有意	有意でない	合計
Glycolysis	9	0	9
その他	80	193	273
合計	89	193	282

となる。Glycolysis の p-value はそれぞれ 1.0030×10^{-7}、2.3266×10^{-5} であり、仮に Table2(右)の Glycolysis で有意でない代謝物が 1 のときは $p=1.6610\times10^{-4}$ となり、いずれにしても全物質を用いた時の方が p 値は小さくなる。これら 2 つの違いは Glycolysis 以外のパスウェイについて有意でない代謝物の数がそれぞれ 563 と 193 であり、これが p 値に大きく影響を及ぼしていることがわかる。

この結果より、検出されなかった代謝物の扱いによって結果が異なることがわかる。Table2(左)の結果より、全ての物質を用いる場合は、検出されなかった物質を有意でない物質とみなしていることがわかる。一方で検出対象物質を用いる場合は、検出されなかった物質は、検出された物質の中で、有意である物質の割合を推定値として用いることとなり、より保守的な推定値となっている。さらに特定の代謝パスウェイ、ここでは Glycolysis に含ま

れない代謝物についても同様に、全ての物質を用いる場合は、検出されなかった物質は全て有意でないとみなしている。しかしながら実際は、検出されなかった物質は統計的に有意でないとみなすのは問題であり、より保守的な検出対象物質を用いる場合の方が現実の設定に近いと考えられる。また特定の代謝パスウェイ以外の有意でない代謝物の数、ここでは563と193の違いが結果に大きく影響を及ぼすことを考えても、検出物質を対象としたORAを行う方が、より保守的な結果が得られることがわかる。

3-2. SSEA の解析結果(1)　S/N を基準として検出された物質群を対象

通常状態で飼育したマウスと、12 時間絶食条件下で飼育したマウスの肝臓のメタボロームデータ[1]を用いて、S/N の閾値として 3 を設定し、検出された物質とそれ以外(特定のサンプルで検出されなかった物質またはパスウェイに含まれているが解析対象ではない物質)での ORA による SSEA を行った。サンプル毎に SSEA スコアを計算し、SSEA スコアを用いて主成分分析を行った結果を図 1 に示す。

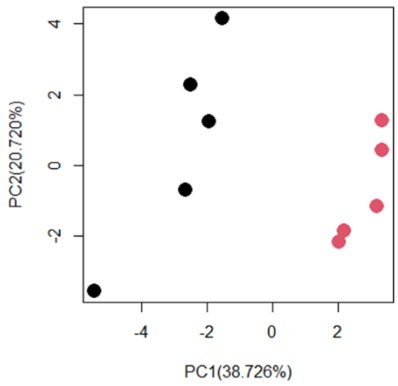

図 1. 絶食マウスでの SSEA スコアに対する主成分分析

第 1 主成分スコアで 2 群間が分かれていることから、第 1 主成分負荷量から 2 群で差のある代謝パスウェイを確認した。第 1 主成分負荷量の p 値に対して Benjamini-Hochberg 法による補正を行って得られた q 値が 0.05 未満の条件を満たす代謝パスウェイは、Glycolysis (R=-0.8984、p=0.0004、q=0.0043)、Alanine, aspartic acid and asparagine metabolism (R=-0.8936、p=0.0005、q=0.0043)、Taurine metabolism (R=-0.8973、p=0.0004、q=0.0043)、Valine, leucine and isoleucine metabolism (R=-0.8601、p=0.0014、q=0.0092)、Pentose phosphate pathway (R=-0.7827、p=0.0074、q=0.0387)となった。従来の方法である、統計的に有意な代謝物を用いて ORA によるエンリッチメント解析を行った結果と比較しても、12 時間の絶食によって、Glycolysis、Pentose phosphate pathway は統計的に有意に低下しており、部分的に同様の結果が得られることが確認された。

3-3. SSEA の解析結果(2)　z-score を基準として低値を示す物質群を対象

物質毎にスケーリングを行って得られた z-score に対して、その閾値を 0 として、正の値を示す代謝物とそれ以外(検出された物質の中で z-score が 0 または負の値を示す物質、特定のサンプルで検出されなかった物質、パスウェイに含まれているが解析対象でない物質)を対象として、ORA による SSEA を行った。これは、z-score が 0 または負の値を示す物質を検出されない物質とみなしていることに相当する。SSEA スコアの値を用いて主成分分析を行った結果を図 2(A)に示す。

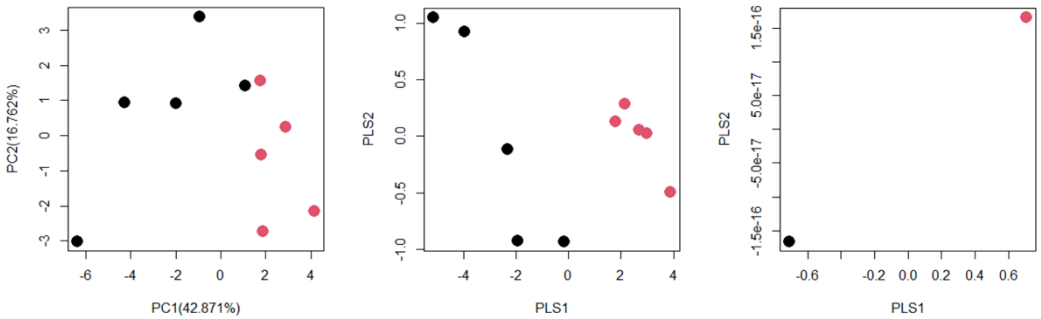

図 2. (A) z-score が負の値を欠損値としたスコアに対する主成分分析の結果
(B) PLS の結果(説明変数)、(C)PLS の結果(目的変数)

第 1 主成分スコア(図 2(A))で通常飼育と 12 時間絶食で群間差が確認されたが、図 1 の結果と比べると群間差は小さい。そこで PLS を行った結果を図 2(B、C)に示す。説明変数の第 1PLS スコア(図 2(B))で群間差が確認されたことから、第 1PLS 負荷量より q<0.05 で有意なパスウェイは、Glycolysis (R=-0.9608、$p=9.88\times10^{-6}$、q=0.00012)、Pentose phosphate pathway(R=-0.9576、$p=1.34\times10^{-5}$、q=0.00012)、Histidine metabolism(R=0.9650、$p=6.32\times10^{-6}$、q=0.00012)であった。このうち、Glycolysis と Pentose phosphate pathway は負の値になっており、目的変数の第 1PLS スコアの結果(図 2(C))からこれらの代謝パスウェイは 12 時間絶食で低下していることを示しており、これまでの結果とも一致している。一方、Histidine metabolism に含まれる z-score を基準として得られた高い値を示す代謝物群が、12 時間の絶食で増加することを示している。

3-4. SSEA の解析結果(2)　プロテオミクスの例

ここまではメタボロミクスの場合について解析結果を紹介してきたが、以下でプロテオミクスの例について紹介する。データは、COVID-19 の血清サンプルのプロテオームデータであり、健常、COVID-19 軽症、重症の順序のある 3 群データを用いた。このデータに対してサンプル毎に値が 0 より大きいタンパク質とそれ以外(特定のサンプルで検出されなかったタンパク質またはパスウェイに含まれているが解析対象ではないタンパク質)で SSEA を行い、得られた SSEA スコアに対して主成分分析を行った結果を図 3 に示す。

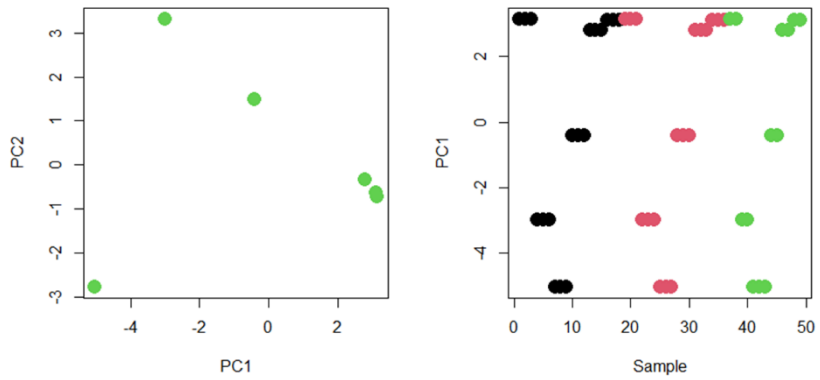

図 3. COVID-19 での SSEA スコアに対する主成分分析

この結果は、COVID-19 のデータが 5 か 6 つのパターンに集約され重ねて描かれており、群間差を確認することが出来なかった。これはそれぞれのサンプルで検出されるタンパク質が共通しており、その結果このような少ないパターンの SSEA スコアに集約された結果、図 3 のような主成分分析の結果となったと考えられる。

そこで、3-3 と同様に z-score が正の値を示すタンパク質とそれ以外(検出されたタンパク質の中で z-score が 0 または負の値を示す物質、特定のサンプルで検出されなかったタンパク質、パスウェイに含まれているが解析対象でないタンパク質)で ORA による SSEA を行い、SSEA スコアに対して群に順序のある PLS である PLS-ROG を行った結果を図 4 に示す。

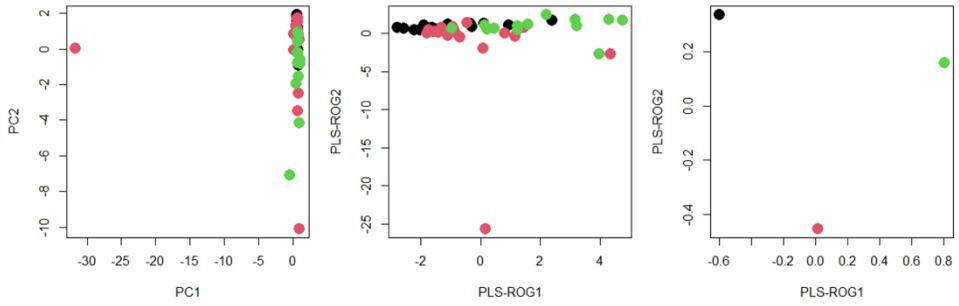

図 4. (A)COVID-19 での SSEA スコアに対する主成分分析
(B)PLS-ROG(説明変数)、(C)PLS-ROG(目的変数)

PLS-ROG の第 1 成分で健常、COVID-19 の軽症、重症の順に並んでいることから、第 1 成分が重症度と関連していると考えられる。そこで第 1PLS-ROG 負荷量について $q<0.05$ のパスウェイは、Alternative Complement Pathway ($R=0.5985$、$p=5.58\times10^{-6}$、$q=0.00022$)、Lectin-Induced Complement Pathway ($R=0.5368$、$p=7.01\times10^{-5}$、$q=0.0014$)、Complement

Pathway (R=0.5133、p=0.00016、q=0.0021)、Classical Complement Pathway (R=0.4771、p=0.00053、q=0.0052)となった。これらはいずれも免疫と関連しており、また COVID-19 の重症度と免疫系との関連も様々報告されている。

3-5. SSEA の解析結果(2)　マルチオミックスの例

ORA による SSEA スコアをプロテオームとメタボロームそれぞれについて計算し、その結果を統合することでマルチオミックス解析に利用することも出来る。ここでは COVID-19 のプロテオーム、メタボロームそれぞれ ORA による SSEA を行った。得られた SSEA スコアを用いて、相関解析を行った結果を図 5 に示す。

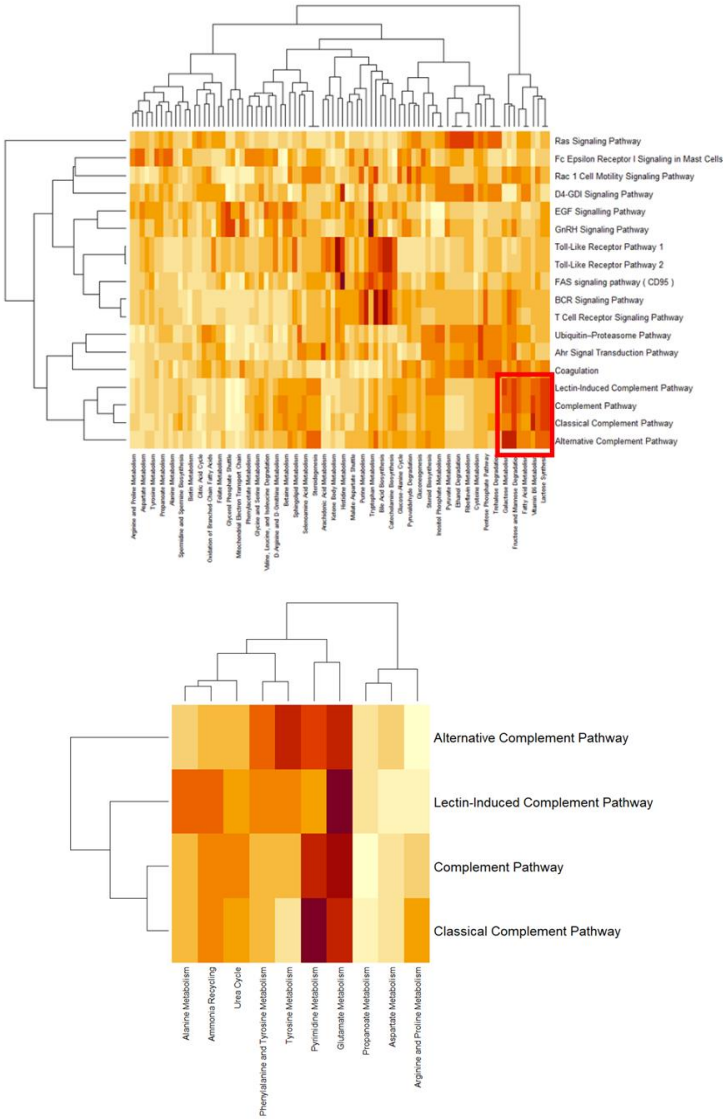

図 5. (A) プロテオームとメタボロームに対する SSEA スコアの相関係数の Heatmap
(B) 相関の高い組み合わせの Heatmap

プロテオームとメタボロームの相関係数のクラスタリングの結果より、4 つのプロテオームから得られた Complement Pathways（Alternative Complement Pathway、Lectin-Induced Complement Pathway、Complement Pathway、Classical Complement Pathway）と相関の高いメタボロームから得られた代謝パスウェイを含むクラスターを抽出することが出来た。この部分に着目すると、特に Pyrimidine Metabolism と Glutamate Metabolism が 4 つの Complement Pathways と相関が高いことが確認された。これらはさらなる検討が必要となるものの、COVID-19 の重症度と関連している代謝パスウェイであることが示唆された。

Conclusion

本研究では、シングルサンプルエンリッチメント解析を絶食マウスのメタボロームデータと COVID-19 のプロテオームデータに適用した。解析結果より、従来の解析と部分的に同様の結果が得られることが確認された。また本手法は、検出された化合物とそれ以外の化合物の情報を利用している点で、従来とは異なるエンリッチメント解析手法であり、全ての化合物を完全に網羅することが難しく欠損値が多い質量分析を用いたオミックスデータに対して、特に有用であると考えられる。また SSEA スコアを用いた解析は非常に柔軟に利用でき、本研究で示した多変量解析、マルチオミックス解析だけでなく、機械学習など様々な解析にも利用できることから、今後さらに利用が拡大することが期待される。

References

[1] Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA, Global function profiling of gene expression. *Genomics*. 2003, 81: 98-104.

[2] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005, 102 (43): 15545-15550.

[3] Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* **14**, 7 (2013).

[4] Wieder, C., Lai, R.P.J. & Ebbels, T.M.D. Single sample pathway analysis in metabolomics: performance evaluation and application. *BMC Bioinformatics* 23, 481 (2022).

[5] Yamamoto, H., Fujimori, T., Sato, H. *et al.* Statistical hypothesis testing of factor loading in principal component analysis and its application to metabolite set enrichment analysis. *BMC Bioinformatics* **15**, 51 (2014).

[6] B. Shen, et al, Proteomic and Metabolomic Characterization of COVID-19 Patient Sera, *Cell*. 182 (2020) 59-72.e15.

Supplementary information

3-3 では、z-score が正の値を検出できた代謝物とみなして ORA による SSEA を行ったが、z-score が負の値に着目して ORA による SSEA を行うことも出来る。ただしこの場合は、検出出来た物質とそれ以外の物質の比較という本研究で提案するコンセプトとは異なることから、参考として結果を紹介する。

3-3 と同様に、z-score が負の値を示す代謝物とそれ以外(検出された物質の中で z-score が 0 または正の値を示す物質、特定のサンプルで検出されなかった物質、パスウェイに含まれているが解析対象でない物質)を対象として、ORA による SSEA を行った。SSEA スコアの値を用いて主成分分析を行った結果を図 3(A)に示す。

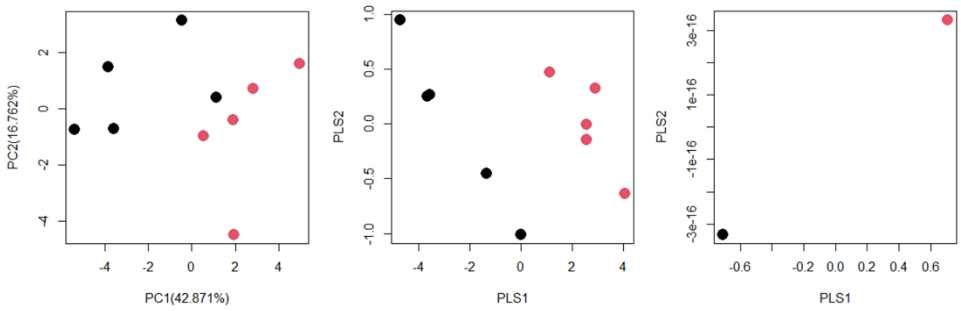

図 6. (A) z-score が正の値を欠損値としたスコアに対する主成分分析の結果
(B) PLS の結果(説明変数)、(C)PLS の結果(目的変数)

上の例と同様に、第 1 主成分スコア(図 3(A))で通常飼育と 12 時間絶食で部分的に群間差が確認されたが、図 1 の結果と比べると群間差は小さい。そこで PLS を行った結果を図 3(B、C)に示す。説明変数の第 1PLS スコア(図 2(B))で群間差が確認されたことから、第 1PLS 負荷量より q<0.05 で有意なパスウェイは、Glycolysis(R=0.9892, p=5.839×10^{-8}、q=1.518×10^{-6})、Pentose phosphate pathway(R=0.8579、p=0.0015、q=0.0375)、TCA cycle (R=0.8567、p=0.0015、q=0.0375)であった。ただし PLS 負荷量の値は正なので、12 時間絶食では低値を示す。これは、低値を示す代謝物を選び出し ORA を行っているので、正負が逆転することに起因している。これらの結果より、本結果は 12 時間絶食で低下していることを示しており、特にパスウェイ情報をそのまま用いて ORA によるエンリッチメント解析を行った場合の結果(Table1)と良く一致していることが確認された。