Multiset partial least squares with rank order of groups for integrating multi-omics data

Hiroyuki Yamamoto *

Human Metabolome Technologies, Inc., 246-2 Mizukami, Kakuganji, Tsuruoka, Yamagata 997-0052, Japan

*Corresponding author h.yama2396@gmail.com

Abstract

As more multi-omics data such as metabolome, proteome and transcriptome data are acquired, statistical analyses to integrate multi-omics data has been required. Multivariate analysis for integration of multi-omics data has been proposed such as multiset partial least squares (PLS). However, when we want to extract scores about the rank order of groups such as severity of disease as well as difference of groups, we could not always extract scores with rank order of groups by using multiset PLS. Therefore, we proposed multiset PLS with rank order of groups (multiset PLS-ROG) to integrate multiple omics dataset. We visualized multiple omics data by using multiset PLS-ROG in two studies of multi-omics and multi-organ derived metabolome data. After we visualize data and focus on a specific score with phenotype of interest, it is important to select compounds to identify biomarker candidates or make biological inferences. In ordinary principal component analysis (PCA) or PLS, we could select statistical significantly compounds correlated with PC or PLS score by using PC or PLS loadings. However, multiset PLS loading has not been used because its statistical property has not been clarified. Then, we clarified statistical property of multiset PLS and multiset PLS-ROG loading, and we could identify statistically significant compounds by using statistical hypothesis testing. We implemented multiset PLS-ROG to R loadings package, freely available in CRAN.

Keywords: partial least squares, rank order of groups, multi-omics, metabolomics, COVID-19

1. Introduction

Recently, many omics studies such as transcriptome, proteome and metabolome and integrated analysis for these multi-omics data have been reported [1, 2]. In an early study of multi-omics, Ishii et al [3] visualized multi-omics data by mapping gene expression, proteome, and metabolome data of *E. coli* onto a single metabolic pathway. This approach has extended to visualize relationships between each omics layer by connecting the regulatory relationships as trans-omics [4]. Many other multi-omics studies such as integration of metagenome and metabolome for gut microbiota [5] have been actively reported. Multi-omics studies were also not only applied for making biological inference in experimental biology, but also for genome-wide association between genome and protein or metabolite such as quantitative trait in clinical research [6].

For statistical analysis of omics data, various univariate analysis methods such as *t*-test and multivariate analysis methods have been applied. However, a standard statistical method for integrating multi-omics data has not yet been established, compared to the routine application of certain statistical analyses such as hierarchical clustering in transcriptomics and principal component analysis (PCA) in metabolomics. In the context of the necessity for standard statistical analysis methods for integrating multi-omics data, multiset multivariate analysis has been applied since the early 2010 to integrate multi-

omics data [7]. One of the most representative tools of multiset multivariate analysis for integrating multi-omics data is mixOmics [8] of the Bioconductor R package that has been cited by above one thousand papers in 2022. The mixOmics is one of the standard statistical analysis tools for integrating multi-omics data.

In mixOmics package, multi-set partial least squares discriminant analysis (PLS-DA) is mainly adopted. While variable importance in projection (VIP) has often been used to select important compounds (i.e., genes, proteins and metabolites) in PLS-DA, correlation coefficient between multiset PLS-DA scores and each compound levels is used to select important compounds in multiset PLS-DA of mixOmics package. On the other hands, PC or PLS loadings have been also used to select important metabolites in metabolomics. We previously clarified the statistical properties of the PC or PLS loadings and applied statistical hypothesis testing of PC [9] or PLS loading [10] to select statistically significant metabolites. As well as PCA or PLS, the statistical properties of multiset PLS loadings are expected to be clarified because we can select statistically significant compounds by using statistical hypothesis testing of multiset PLS loading.

In practical data analysis, we have sometimes an additional information about groups such as the severity of the disease or the concentration of the drug to be administered. When we applied PLS to this type of data, the rank order of the groups is not necessarily

represented in the PLS scores. Therefore, we previously proposed PLS with rank order of groups (PLS-ROG). In this study, we extended PLS-ROG to multiset PLS-ROG for integrating multiple omics dataset. We not only organized multiset PLS-ROG from a theoretical perspective but also applied multiset PLS-ROG to multi-omics dataset. As a practical application of multiset PLS-ROG, we reanalyzed the multi-omics data of proteome and metabolome data from human serum samples on severity of COVID-19 and reanalyzed the multiple metabolome data from rabbit samples derived from different organs and plasma on the effects of drugs on hyperlipidemia. From practical data analysis in two studies, we confirmed the effects of the parameters on multiset PLS-ROG scores and on the change of highly ranked statistically significant compounds. Multiset PLS-ROG also has the advantage that is easy to implement, and can be computed using loadings package of R, which is freely available from the CRAN website.

2. Methods

2-1. Multiset PLS-ROG

PLS is formulated as maximization of covariance between score for explanatory variable (i.e., omics data) and for response variable (i.e., group information) [10,11]. And

PLS-ROG [10] is formulated as well as PLS, but differs from PLS in that a smoothing term for the group-wise mean of the scores for response variable is added to the constraint condition of PLS. Multiset PLS is also formulated as maximization of sum of covariance between scores for all combinations of each explanatory variables and between score for response and each explanatory variable. As well as multiset PLS, multiset PLS-ROG is formulated as maximization of sum of covariance and almost the same constraint condition as PLS-ROG as follows,

$$\begin{gathered}\max \sum_{i=1}^{K-1} \sum_{j \neq i}^{K} \tau_{i,j} \mathrm{cov}(\mathbf{t_i}, \mathbf{t_j}) + \sum_{k=1}^{K} \tau_{k,Y} \mathrm{cov}(\mathbf{t_k}, \mathbf{s}) \\ s.t. \sum_{k=1}^{K} \mathbf{w_{x,k}}' \mathbf{w_{x,k}} + (1-\kappa) \mathbf{w_y}' \mathbf{w_y} + \kappa \mathbf{w_y}' \mathbf{Y}' \mathbf{P}' \mathbf{D}' \mathbf{D} \mathbf{P} \mathbf{Y} \mathbf{w_y}' = 1.\end{gathered} \tag{1}$$

K is the number of datasets. The parameter $\tau_{i,j}$ represents the strength of the connection between i-th and j-th omics dataset and $\tau_{k,Y}$ represents the strength of the connection between k-th dataset and the response variable. These parameters $\tau_{i,j}$ and $\tau_{k,Y}$ are set so that these values sum up to 1 as follows,

$$\sum_{i=1}^{K-1} \sum_{j \neq i}^{K} \tau_{i,j} + \sum_{k=1}^{K} \tau_{k,Y} = 1.$$

In the implementation, we summarized $\tau_{i,j}$ and $\tau_{k,Y}$ as the following matrix $\boldsymbol{\tau}$,

$$\boldsymbol{\tau} = \begin{bmatrix} 0 & \tau_{1,2} & \cdots & \tau_{1,K} & \tau_{1,Y} \\ \tau_{2,1} & 0 & \cdots & \tau_{2,K} & \tau_{2,Y} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \tau_{K,1} & \tau_{K,2} & \cdots & 0 & \tau_{K,Y} \\ \tau_{Y,1} & \tau_{Y,2} & \cdots & \tau_{Y,K} & 0 \end{bmatrix}. \quad (2)$$

The $\boldsymbol{\tau}$ is a symmetric matrix whose diagonal elements are zero because the same dataset is not connected to each other. The cov($\mathbf{t}_k$, $\mathbf{s}$) means the covariance between the k-th score for explanatory variable $\mathbf{t}_k=\mathbf{X}_k\mathbf{w}_{x,k}$ and the score for response variable $\mathbf{s}=\mathbf{Y}\mathbf{w}_y$. The $\mathbf{w}_{x,k}$ is the weight vector for the k-th data matrix $\mathbf{X}_k$ and $\mathbf{w}_y$ is the weight vector for the dummy matrix corresponds to the response variable. The k-th data matrix $\mathbf{X}_k$ includes samples in rows and metabolites in column, and the matrix **Y** includes samples in row and group number in column. For example, when the number of groups is g, the matrix **Y** is set as follows,

$$\mathbf{Y} = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0}_{n_g} & \mathbf{0}_{n_g} & \cdots & \mathbf{1}_{n_g} \end{bmatrix}.$$

The elements of the matrix **Y** are vectors, for example, $\mathbf{1}_{n_1}$ means an n_1-dimensional vector with 1 as an element and $\mathbf{0}_{n_1}$ means an n_1-dimensional vector with 0 as an element. In the practical calculations, **X** and **Y** were mean-centered matrix, and the matrix **X** was further divided by the standard deviation for each metabolite levels (i.e., autoscaling) in this study. The first order differential matrix **D** and averaging matrix **P** are set for the g-class classification problems as follows,

$$\mathbf{D} = \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

and

$$\mathbf{P} = \begin{bmatrix} (1/n_1)\,\mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & (1/n_2)\,\mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_g} & \mathbf{0}_{n_g} & \cdots & (1/n_g)\,\mathbf{1}_{n_g} \end{bmatrix}.$$

κ is a smoothing parameter with $0 \leq \kappa < 1$. When κ is closer to 1, the rank order of groups is more likely to appear in the multiset PLS-ROG scores because the effect of smoothing term of $\kappa \mathbf{w_y}'\mathbf{Y}'\mathbf{P}'\mathbf{D}'\mathbf{DPYw_y}'$ is relatively large in eq. (1). When κ equals to 0, multiset PLS-ROG equals to the multiset PLS.

The maximization problem under constraint condition in eq. (1) can be written by using the method of Lagrange multiplier as follows,

$$J = \frac{1}{N}\sum_{i=1}^{K-1}\sum_{j\neq i}^{K} \tau_{i,j}\mathbf{w_{x,i}}'\mathbf{X_i}'\mathbf{X_j}\mathbf{w_{x,j}} + \frac{1}{N}\tau_{k,Y}\sum_{k=1}^{K}\mathbf{w_{x,k}}'\mathbf{X_k}'\mathbf{Y}\mathbf{w_y}$$

$$+ \boldsymbol{\lambda}\left\{1 - \left(\sum_{k=1}^{K}\mathbf{w_{x,k}}'\mathbf{w_{x,k}}\right) - (1-\kappa)\mathbf{w_y}'\mathbf{w_y} - \kappa\mathbf{w_y}'\mathbf{Y}'\mathbf{P}'\mathbf{D}'\mathbf{DPYw_y}\right\}.$$

This equation is partially differentiated by the weight vector of $\mathbf{w_{x,1}}, \mathbf{w_{x,2}}, \ldots, \mathbf{w_{x,K}}$ and $\mathbf{w_y}$ respectively, and setting it to 0.

$$\frac{\partial J}{\partial \mathbf{w_{x,k}}} = \frac{1}{N}\sum_{j\neq k}^{K}\tau_{k,j}\mathbf{X_k}'\mathbf{X_j}\mathbf{w_{x,j}} + \frac{1}{N}\tau_{k,Y}\mathbf{X_k}'\mathbf{Y}\mathbf{w_y} - 2\lambda\mathbf{w_{x,k}} = 0 \qquad (3)$$

$$\frac{\partial J}{\partial \mathbf{w_y}} = \frac{1}{N}\sum_{k=1}^{K} \tau_{k,Y}\mathbf{Y'X_k w_{x,k}} - 2\lambda\{(1-\kappa)\mathbf{I} + \kappa\mathbf{Y'P'D'DPY}\}\mathbf{w_y} = 0.$$

where k=1, 2, ... , K and **I** is the identity matrix. Here, we review the meaning of λ. We multiplied by $\mathbf{w_y}$ from the left of the lower expression in eq. (3),

$$\frac{1}{N}\sum_{k=1}^{K} \tau_{k,Y}\boldsymbol{w_y}'\mathbf{Y'X_k w_{x,k}} = 2\lambda\boldsymbol{w_y}'\{(1-\kappa)\mathbf{I} + \kappa\mathbf{Y'P'D'DPY}\}\boldsymbol{w_y}.$$

The left side of the above equation is a weighted sum of covariance between the score for explanatory $\mathbf{t_k}$ and response variable **s**, and λ can be written as follows,

$$\lambda = \frac{1}{2}\sum_{k=1}^{K} \tau_{k,Y} cov(\mathbf{s}, \mathbf{t_k}).$$

Finally, the eq. (3) can be written as the following generalized eigenvalue problem,

$$\begin{bmatrix} \mathbf{0} & \frac{1}{N}\tau_{1,2}\mathbf{X_1}'\mathbf{X_2} & \cdots & \frac{1}{N}\tau_{1,K}\mathbf{X_1}'\mathbf{X_K} & \frac{1}{N}\tau_{1,Y}\mathbf{X_1}'\mathbf{Y} \\ \frac{1}{N}\tau_{2,1}\mathbf{X_2}'\mathbf{X_1} & \mathbf{0} & \cdots & \frac{1}{N}\tau_{2,K}\mathbf{X_2}'\mathbf{X_K} & \frac{1}{N}\tau_{2,Y}\mathbf{X_2}'\mathbf{Y} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{1}{N}\tau_{K,1}\mathbf{X_K}'\mathbf{X_1} & \frac{1}{N}\tau_{K,2}\mathbf{X_K}'\mathbf{X_2} & \cdots & \mathbf{0} & \frac{1}{N}\tau_{K,Y}\mathbf{X_K}'\mathbf{Y} \\ \frac{1}{N}\tau_{1,Y}\mathbf{Y'X_1} & \frac{1}{N}\tau_{2,Y}\mathbf{Y'X_2} & \cdots & \frac{1}{N}\tau_{K,Y}\mathbf{Y'X_K} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{w_{x,1}} \\ \mathbf{w_{x,2}} \\ \vdots \\ \mathbf{w_{x,k}} \\ \mathbf{w_y} \end{bmatrix} = \lambda \begin{bmatrix} 2\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 2\mathbf{I} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & 2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & 2\{(1-\kappa)\mathbf{I} + \kappa\mathbf{Y'P'D'DPY}\} \end{bmatrix} \begin{bmatrix} \mathbf{w_{x,1}} \\ \mathbf{w_{x,2}} \\ \vdots \\ \mathbf{w_{x,k}} \\ \mathbf{w_y} \end{bmatrix}. \tag{4}$$

We focus on the component with the positively larger eigenvalue. For example, the component corresponds to the largest positive eigenvalue is the first component, and the component corresponds to the next largest positive eigenvalue is the second component.

The corresponding eigenvector is the weight vector **w** for each component. And, the multiset PLS-ROG score for the k-th dataset and the score for the response variable can be calculated with $\mathbf{t_k}=\mathbf{X_k w_{x,k}}$ and $\mathbf{s}=\mathbf{Y w_y}$.

So far, we have considered the case where we connected between explanatory variables for each dataset, and between the response variable and explanatory variable in each dataset (Fig.1(a)). We also consider the case where we don't connect between explanatory variables for each dataset (Fig.1 (b)). If the connection between the explanatory variables for each datasets were not made, $\tau_{i,j}$ equals to 0 in equation (1). This eliminates the term $\sum_{i=1}^{K-1}\sum_{j\neq i}^{K}\tau_{i,j}\mathrm{cov}(\mathbf{t_i},\mathbf{t_j})$ in eq (1) and can be written as follows,

$$\begin{gathered}\max\sum_{k=1}^{K}\tau_k\mathrm{cov}(\mathbf{t}_k,\mathbf{s})\\ s.t.\sum_{k=1}^{K}(1-\kappa)\mathbf{w_{x,k}}'\mathbf{w_{x,k}}+\kappa\mathbf{w_y}'\mathbf{Y}'\mathbf{P}'\mathbf{D}'\mathbf{DPYw_y}'=1\end{gathered}\tag{5}$$

When we set τ_k as 1, the summation of covariance in eq. (5) can be re-written as covariance of the summation as follows,

$$\sum_{k=1}^{K}\mathrm{cov}(\mathbf{t}_k,\mathbf{s})=\mathrm{cov}\left(\sum_{k=1}^{K}\mathbf{t}_k,\mathbf{s}\right)=\mathrm{cov}(\mathbf{t},\mathbf{s}).\tag{6}$$

In the eq. (6), covariance between sum of the scores $\mathbf{t_k}$ for the explanatory variables in each dataset k and the score for response variable is equivalent to the covariance between the score **t**, combining each dataset in the column direction into a single data matrix, and

the score for response variable. In this case, we can compute multiset PLS-ROG by using PLS-ROG for this combined data matrix when samples in each row and compounds in each column. Multiset PLS-ROG score for explanatory variables in each dataset k can be computed as $\mathbf{t_k}=\mathbf{X_k}\mathbf{w_{x,k}}$ by using the divided weight vector $\mathbf{w_{x,k}}$ and each dataset $\mathbf{X_k}$ (k=1,2,…,K).

In this study, we developed multiset PLS-ROG in supervised case and explained the computational framework for multiple datasets. Additionally, multiset PLS in unsupervised case, that connected between omics datasets only, can be calculated in the same framework.

2-2. Statistical property of multiset PLS-ROG loading

We visualize omics data with multiset PLS-ROG score and focus on that score associated with the phenotypes such as group information, and then select important compounds corresponding to this score. We can select relative important compounds by using the multiset PLS-ROG weight vector $\mathbf{w_x}$ in eq (4). However, we cannot set a specific threshold for multiset PLS-ROG weight and select important compounds by using a statistical criterion because multiset PLS-ROG weight is set for computational convenience but not statistically meaning. In PCA and PLS, we clarified statistical

properties of PC weight or PLS weight and were able to select statistically significant compounds by using statistical hypothesis testing of PC or PLS loading. As with PCA and PLS, it is important to clarify the statistical properties of the multiset PLS-ROG weight in order to select significant compounds by using statistical criteria.

To clarify the statistical property of multiset PLS-ROG loading, we first consider the weighted sum of covariance $C_{x,k,p}$ between the scores for the explanatory and response variables and the p-th metabolite in the k-th dataset.

$$\begin{aligned} C_{x,k,p} &= \sum_{j\neq k}^{K} \tau_{k,j} cov\left(\mathbf{t_j}, \mathbf{x}_{\boldsymbol{k},\mathbf{p}}\right) + \tau_{k,Y} cov\left(\mathbf{s}, \mathbf{x}_{\boldsymbol{k},\mathbf{p}}\right) \\ &= \sum_{j\neq k}^{K} \tau_{k,j} \frac{1}{N} \mathbf{w_j}'\mathbf{X_j}'\mathbf{X_{k,p}} + \tau_{k,Y} \frac{1}{N} \mathbf{w_y}'\mathbf{Y}'\mathbf{X_{k,p}}. \end{aligned} \tag{7}$$

We plugged in the equation, $\frac{1}{N}\sum_{j\neq k}^{K} \tau_{k,j}\mathbf{X_k}'\mathbf{X_j}\mathbf{w_{x,j}} + \frac{1}{N}\tau_{k,Y}\mathbf{X_k}'\mathbf{Y}\mathbf{w_y} = 2\lambda\mathbf{w_{x,k}}$, that is the transposed equation of a part of the eq. (3), into the eq. (7). Finally, the eq. (7) can be written as follows,

$$C_{x,k,p} = \left\{\frac{1}{N}\sum_{j\neq k}^{K} \tau_{k,j}\mathbf{w_j}'\mathbf{X_j}'\mathbf{X_k} + \frac{1}{N}\tau_{k,Y}\mathbf{w_y}'\mathbf{Y}'\mathbf{X_k}\right\}\mathbf{c} = \left(2\lambda\mathbf{w_{x,k}}\right)'\mathbf{c} = 2\lambda w_{x,k,p}, \tag{8}$$

where **c** is the column vector in which the *p*-th element is 1 and the others are 0. From the eq. (8), we were able to confirm that multiset PLS-ROG weight $\mathbf{w_{x,k}}$ is proportional to the weighted sum of correlation coefficient between the score for explanatory variable $\mathbf{t_j}$ and *p*-th variable in k-th dataset $\mathbf{x_{k,p}}$, and between response variable **s** and $\mathbf{x_{k,p}}$. And using

the eq. (6), the weighted sum of covariance $C_{x,k,p}$ can be written as follows,

$$C_{x,k,p} = \sum_{j \neq k}^{K} \tau_{k,j} cov(\mathbf{t_j}, \mathbf{x}_{\boldsymbol{k},\mathbf{p}}) + \tau_{k,Y} cov(\mathbf{s}, \mathbf{x}_{\boldsymbol{k},\mathbf{p}})$$
$$= cov\left(\sum_{j \neq k}^{K} \tau_{k,j} \mathbf{t_j} + \tau_{k,Y} \mathbf{s}, \mathbf{x}_{\boldsymbol{k},\mathbf{p}}\right). \quad (9)$$

Fron the eq. (9), the weighted sum of covariance $C_{x,k,p}$ equals to the covariance between the weighted sum of score for explanatory **t** and response variable **s**. Based on these results, we defined the multiset PLS-ROG loading $R_{x,k,p}$ as follows,

$$R_{x,k,p} = corr(\boldsymbol{u_k}, \boldsymbol{x})$$
$$= \frac{cov\left(\sum_{j \neq k}^{K} \tau_{k,j} \mathbf{t_j} + \tau_{k,Y} \mathbf{s}, \mathbf{x_p}\right)}{\sqrt{var\left(\sum_{j \neq k}^{K} \tau_{k,j} \mathbf{t_j} + \tau_{k,Y} \mathbf{s}\right)} \sqrt{var(\mathbf{x_p})}}. \quad (10)$$

When each compound levels $\mathbf{x_{k,p}}$ of p-th compound in k-th dataset are centered and scaled to unit variance (i.e., autoscaling), the var($\mathbf{x_{k,p}}$), variance of $\mathbf{x_{k,p}}$, is 1. And, the variance of the weighted sum of scores can be regarded as a constant because the weighted sum of scores does not depend on the p-th compound of $x_{k,p.}$ Therefore, the relationship between the multiset PLS-ROG loading $R_{x,k,p}$ and the weighted vector $\mathbf{w_{x,k}}$ also can be written as follows,

$$R_{x,k,p} = \frac{2\lambda w_{x,k,p}}{\sqrt{var\left(\sum_{j \neq k}^{K} \tau_{k,j} \mathbf{t_j} + \tau_{k,Y} \mathbf{s}\right)}}. \quad (11)$$

In our previous study of PCA, PLS and PLS-ROG, *p*-values of loadings can be calculated by using *t*-distribution because these loadings can be defined as correlation coefficient

itself. As well as these approaches, the statistical hypothesis testing of factor loading in the multiset PLS and PLS-ROG can be performed by using that $\mathrm{t-statistics} = R_{x,k,p}\sqrt{N-2}/\sqrt{1-{R_{x,k,p}}^2}$ follows a *t*-distribution.

2-3. Implementation and computation

All statistical analysis in this study was computed by using R ver. 4.3.3, and we implemented the multiset PLS and multiset PLS-ROG to R loadings package. The loadings package is freely available in CRAN web site [https://cran.r-project.org/web/packages/loadings/]. We computed generalized eigenvalue problem by using geigen package. The *q*-value of multiset PLS-ROG loading was computed by Benjamini-Hochberg method.

3. Results

3-1. Multi-omics data in COVID-19 study

We re-analyzed multi-omics data of proteome and metabolome in COVID-19 study by Shen et. al. [11] as an application of multiset PLS-ROG. This dataset includes three groups of healthy control, mild and severe COVD-19. Our goal is to integrate and visualize multi-omics data and find score associated with severity of COVID-19, and then

find compounds (proteins or metabolites) associated with the severity of COVID-19. We performed multiset PLS-ROG with the parameter κ as 0.999, default parameter of PLS-ROG in loadings packages, and with the weight of connectivity between group and each datasets are set to the same value as $\tau_{1,2}=\tau_{1,Y}=\tau_{2,Y}=1/3$.

Multiset PLS-ROG score for explanatory variables in the proteome (Fig.2 (A_1)) and metabolome (Fig.2 (B_1)) datasets showed that the specific samples of healthy subjects (JBDZ2, JBDZ3, and JBDZ6) located at a distance from many other samples in first multiset PLS-ROG score, and these three samples might be considered as outliers. We also confirmed three outlier samples in first and second score by unsupervised multiset PLS (Fig.S1) and supervised multiset PLS (Fig.S2). These results suggest that there is a strong association between proteome and metabolome data for specific samples. We also could confirm the rank order of groups, i.e., the severity of COVID-19, in first (Fig.2 (C_2)) and second (Fig.2 (C_3)) multiset PLS-ROG score for response variable. However, group differences and rank order of groups were not found in first multiset PLS-ROG scores of explanatory variables and rank order of groups was found slightly in second multiset PLS-ROG score of explanatory variables, but not clearly. This trend showed similar in multiset PLS scores (Fig S2).

One possible reason why group differences and rank order of groups do not appear

clearly in multiset PLS scores of explanatory variables is that the relationship between groups and each omics data is weak. Then, we confirmed the correlation coefficient between first multiset PLS-ROG score for response variable and for explanatory variable. The correlation coefficient between first multiset PLS-ROG scores for proteome and metabolome data is 0.835, and the correlation coefficient between group and proteome data is 0.35 and the correlation coefficient between group and metabolome data is 0.427, respectively when $\tau_{1,2}=\tau_{1,Y}=\tau_{2,Y}=1/3$. This result showed that the correlation between multiset PLS-ROG score for response and explanatory variables was small, and as a result, the rank order of groups was not shown in multiset PLS-ROG score for explanatory variables. Moreover, we showed the change of correlation coefficient between group and each omics dataset when the weight of connectivity $\tau_{1,Y}$ and $\tau_{2,Y}$ was changed simultaneously from 0.3 to 0.5 in Fig.3. This result showed that the higher the weight of the connectivity between group and each omics dataset, the higher the correlation coefficient between the first multiset PLS-ROG score for response variable and for explanatory variable, and the gradually lower the correlation coefficient between the first multiset PLS-ROG scores of proteome and metabolome data for explanatory variable.

The results so far suggest that not only smoothing parameter κ but also the weight of connectivity τ greatly affect to score in multiset PLS-ROG. Then, we performed multiset

PLS-ROG when the weight of connectivity between group and each omics datasets was set as $\tau_{1,Y}=\tau_{2,Y}=0.48$ (Fig.S3), 0.49 (Fig.S4) or 0.50 (Fig.S5) in which the correlation coefficient changed significantly (Fig.3). In Fig.4, we showed the first multiset PLS-ROG score for explanatory variable of proteome data at $\tau_{1,Y}=\tau_{2,Y}=0.48$, 0.49 or 0.50 to clearly represent how the score changes with the parameter τ. This result showed that the effect of outliers becomes smaller and the rank order of the groups becomes clearer in first multiset PLS-ROG score as the parameter $\tau_{1,Y}$ and $\tau_{2,Y}$ was increased from 0.48 to 0.50. And we picked up the top 3 highly correlated proteins with multiset PLS-ROG score by using multiset PLS-ROG loading. The first, second and third rank proteins were P12111 (R=-0.7966, $p=7.645 \times 10^{-12}$, $q=6.048 \times 10^{-9}$), P13727 (R=-0.7250, $p=3.839 \times 10^{-9}$, $q=1.504 \times 10^{-6}$) and P24821 (R=-0.7195, $p=5.704 \times 10^{-9}$, $q=1.504 \times 10^{-6}$) at $\tau=0.48$, and were Q14520 (R=0.6648, $p=1.889 \times 10^{-7}$, $q=1.494 \times 10^{-4}$), Q15113 (R=-0.6374, $p=8.422 \times 10^{-7}$, $q=2.350 \times 10^{-4}$) and P12111 (R=-0.6363, $p=8.913 \times 10^{-7}$, $q=2.350 \times 10^{-4}$) at $\tau=0.49$ and were Q14520 (R=0.6731, $p=1.167 \times 10^{-7}$, $q=9.233 \times 10^{-5}$), P12259 (R=0.6589, $p=2.636 \times 10^{-7}$, $q=1.042 \times 10^{-4}$) and Q15113 (R=-0.6118, $p=3.007 \times 10^{-6}$, $q=7.929 \times 10^{-4}$) at $\tau=0.50$.

We confirmed outliers in the top 3 ranked proteins of P12111 (Fig5(A)), P13727 (Fig,5(B)) and P24821 (Fig.5(C)) at $\tau_{1,Y}=\tau_{2,Y}=0.48$ but not confirmed rank order of

groups clearly. On the other hands, we confirmed rank order of groups in the top 3 ranked proteins of Q14520 (Fig.5(D)), P12259 (Fig.5(E)) and Q15113 (Fig.5(F)) at $\tau_{1,Y}=\tau_{2,Y}=0.50$ clearly, but outliers were less impressive than top 3 proteins. We could confirm both rank order of groups and (although small impression) outliers in P14520 and Q15113. P14520 and Q15113 ranked top 1 and 2 at $\tau_{1,Y}=\tau_{2,Y}=0.49$, which was between $\tau_{1,Y}=\tau_{2,Y}=0.48$ where the effect of outliers was significant, and $\tau_{1,Y}=\tau_{2,Y}=0.50$ where the effect of rank order groups was significant. These results suggested that, as well as the effect of outliers becomes smaller and the rank order of the groups becomes clearer in first multiset PLS-ROG score as the parameter $\tau_{1,Y}$ and $\tau_{2,Y}$ was increased from 0.48 to 0.50, we could also confirm same tendency in highly ranked proteins identified by multiset PLS-ROG loading as the parameter $\tau_{1,Y}$ and $\tau_{2,Y}$ was increased from 0.48 to 0.50.

In proteome data, the top 3 with the highest correlation proteins in first multiset PLS-ROG loading at $\tau=0.50$ were Q14520 of hyaluronan-binding protein 2 (Fig.5(D)), P12259 of coagulation factor V (Fig.5(F)), and Q15113 of Procollagen C-endopeptidase enhancer 1 (Fig.5(E)). The original paper for this multi-omics data of COVID-19 study discussed the proteins Q14520 and P12259, but not the protein Q15113. One reason that these proteins were not focused on might be that the levels of two proteins of Q14520 and P12259 increased but the level of the protein Q14126 decreased with the severity of

COVID-19. In metabolome data, while we could confirm group difference between healthy control and COVID-19 in first multiset PLS-ROG score ($\tau_{1,Y}=\tau_{2,Y}=0.50$) for explanatory variables in metabolomics, we could not confirm the rank order of groups clearly (Fig.S5 (B_2)). The highest correlated metabolite in first multiset PLS-ROG loading at $\tau_{1,Y}=\tau_{2,Y}=0.50$ was 5-methyluridine (R=-0.7360, $p=1.694\times10^{-9}$, $q=1.580\times10^{-6}$), also known as ribothymidine (Fig.S6). As well as the first multiset PLS-ROG score for explanatory variables, we could confirm the clear difference between healthy control and COVID-19 groups in 5-methyluridine but could not confirm the difference between mild and severe COVID-19, that was, could not confirmed the rank order of groups related with the severity of COVID-19. In the original paper of COVID-19 study [11], 5-methyluridine was not discussed. We inferred the reason was that this metabolite, like protein Q15113, increased but not decreased with the severity of COVID-19.

3-2. Multi-organ metabolome data in hyperlipidemia study

We re-analyzed multi-organ metabolome data of rabbit plasma, liver, heart muscle and brain samples in hyperlipidemia study [13]. This dataset has three groups of wild type, hyperlipidemia rabbit and hyperlipidemia rabbit with drug administration, with three samples in each group. The main purpose of this hyperlipidemia study is to visualize the

effects of drugs on various organs and plasma in hyperlipidemic rabbits. Specifically, we want to find multiset PLS-ROG score that show the pattern that the hyperlipidemia group is approaching to wild type by drug administration, i.e. the pattern showing rank order of wild type, hyperlipidemia with drug administration and hyperlipidemia groups. After we find the multiset PLS-ROG score showing desirable pattern of rank order of groups, we can identify metabolites associated with drug effects by the corresponding multiset PLS-ROG loadings.

As in the previous section of COVID-19 study, we first performed multiset PLS-ROG with the parameter κ as 0.999 and all weight of connectivity was set to the same value, $\tau_{i,j}=\tau_{i,Y}=\tau_{j,Y} = 0.1$ where $i=1,\cdots,4$ and $j=1,\cdots,4$ and $i\neq j$ (Fig. 6). We could confirm the group difference between hyperlipidemia with drug administration and other groups of wild type and hyperlipidemia group in first multiset PLS-ROG and multiset PLS score (Fig.S7), and could confirm rank order of groups in second multiset PLS-ROG and multiset PLS score of a part of explanatory, liver and heart muscle, and response variables. However, we could not confirm rank order of groups in first PLS-ROG score.

As with the previous section, to clarify the effect of the weight of connectivity to multiset PLS-ROG score, we showed the change in correlation coefficient between group and each omics dataset when the weight of connectivity $\tau_{i,Y}$ and $\tau_{j,Y}$, hereinafter denoted

as τ, was changed simultaneously from 0.1 to 0.25 (Fig.7). In this hyperlipidemia study, we used the mean value of the correlation coefficients between the multiset PLS-ROG scores for explanatory variables because a specific organ was connected to multiple other organs, e.g., plasma was connected to liver, heart muscle and brain. In Fig.7, we could confirm that the weight of connectivity affects greatly to the correlation coefficient between multiset PLS-ROG scores for response and explanatory variables as well as the previous section in COVID-19 study. We could also confirm that the correlation coefficients between second PLS-ROG score for explanatory variables of liver and heart muscle and for response variable were slightly decreased at τ=0.23 in Fig.7 (B). This depends on the fact that the pattern of the rank order of groups of wild type, hyperlipidemia with drug administration and hyperlipidemia groups was swapped from second PLS-ROG score to first PLS-ROG score from τ=0.22 to 0.23. In Fig.7 (B), we also confirmed that the correlation coefficient between second PLS-ROG score for explanatory variables of liver and heart muscle and for response variable are high and similar tendency with a variation of the parameter τ. Because statin is known to inhibit cholesterol production in the liver and have an effect on myocardial infarction, it is reasonable that the effect of statin is more clearly confirmed in the liver and heart muscle.

To confirm the impact of parameter τ on the multiset PLS-ROG in more detail, we

performed multiset PLS-ROG at τ =0.23 (Fig. S9), 0.24 (Fig. S10) and 0.25 (Fig. S11). First multiset PLS-ROG scores at τ=0.23, 0.24 and 0.25 for response and explanatory variables (Fig.8) were shown to clarify the effect on the score with respect to a change in the weight of connectivity. As the parameter τ increased from 0.23 to 0.25, we could confirm the rank order of groups of wild type, hyperlipidemia with drug administration and hyperlipidemia groups in first multiset PLS-ROG score for liver, heart muscle and brain clearly, and we could not confirm the rank order of groups but confirm pattens that hyperlipidemia group was approaching to the hyperlipidemia with drug administration group for plasma. As described above, statin have been known to effect on the liver and muscles, and also known to be effective not only in myocardium but also in brain infarction [14].

We have confirmed how the multiset PLS-ROG scores change with the parameter τ. Now, we focused on the results of multiset PLS-ROG in heart muscle samples because group difference and rank order of groups were more clear than other samples. To confirm how highly ranked metabolites correlated with the multiset PLS-ROG score were changed, we picked the top 3 ranked metabolites by multiset PLS-ROG loading at τ =0.23, 0.24 and 0.25 in heart muscle samples. The top 3 metabolites were carnosine (R=0.9489, $p=9.476\times10^{-5}$, q=0.0153) (Fig.9 (A)), citrulline (R=-0.9230, $p=3.872\times10^{-4}$, q=0.0232)

(Fig.9 (B)) and O-acetylserine (R=0.9489, p=4.330 × 10^{-4} , q=0.0232) (Fig.9 (C)) at τ=0.23, and citrulline (R=-0.8790, p=1.797 × 10^{-3} , q=0.1202) (Fig.9 (A)), N6,N6,N6-trimethyllysine (R=-0.8626, p=2.760 × 10^{-3}, q=0.1202) (Fig.9 (D)) and N,N-dimethylglycine (R=0.8503, p=3.681 × 10^{-3}, q=0.1202) (Fig.9 (E)) at τ=0.24 and N,N-dimethylglycine (R=0.9084, p=6.988 × 10^{-4}, q=0.1125) (Fig.9 (E)), tyramine (R=0.8515, p=3.581 × 10^{-3}, q=0.1987) (Fig.9 (F)) and Gly-Gly (R=-0.8385, p=4.744 × 10^{-3}, q= 0.1987) (Fig.9 (G)) at τ=0.25.

We could not confirm metabolites that showed rank order of groups clearly at τ=0.23. Among them, group difference between wild type and other groups were shown clearly and rank order of groups were shown slightly in citrulline of second ranked metabolite (Fig9 (A)). At τ=0.24, citrulline was ranked first and N,N-dimethylglycine that showed rank order of groups clearly was ranked third. Furthermore, at τ=0.25, N,N-dimethylglycine was ranked first and Gly-Gly of third ranked metabolite also showed rank order of groups. And citrulline was forth ranked at τ=0.25. From these results, we could confirm that rank order of groups was clearer in highly ranked metabolites as τ is increased from 0.23 to 0.25.

About the significant metabolites selected by multiset PLS-ROG, dimethylglycine, Gly-Gly and citrulline were mentioned in the original paper. On the other hand, tyramine

of second ranked metabolites at τ=0.25 that showed rank order of groups slightly was not mentioned. It was reported that tyramine was included in a variety of foods and was a compound that triggers hypertensive attacks [15].

4. Discussion

In this section, we would like to discuss some points to keep in mind for practical data analysis by using multiset PLS-ROG. For visualization of data, it is important to check both multiset PLS-ROG scores for explanatory variable and response variable because the pattern of multiset PLS-ROG scores for the explanatory and response variables sometimes differed each other. For example, in COVID-19 study, we could confirm that multiset PLS-ROG score for response variable showed severity of disease clearly, but that for explanatory variables did not show severity of disease (Fig.2). This becomes more important when we select significant compounds by using multiset PLS-ROG loading. In our previous study of PLS-ROG [10, 16], we could identify significant metabolites with rank order of groups by using PLS-ROG loading, if only the response variable showed rank order of the groups. On the other hands, we could identify compounds with rank order of groups by using multiset PLS-ROG loading only when both the explanatory and response variables showed rank order of groups. The reason we need to confirm both the

explanatory and the response variables can be explained theoretically as follows, whereas PLS-ROG loading is defined as the correlation coefficient between PLS-ROG score for the response variable and each compound, multiset PLS-ROG loading is defined as the weighted sum of correlation coefficient between multiset PLS-ROG score of response variable and each compounds and correlation coefficient between multiset PLS-ROG score for other explanatory variable and each compounds.

We would like to discuss about the parameters of great impact in the multiset PLS-ROG to extract scores with rank order of groups. In our previous study of PLS-ROG, as the smoothing parameter κ increased, the rank order of groups in PLS-ROG score became clearer. On the other hand, in multiset PLS-ROG, not only the smoothing parameter κ increased but also the weight of connectivity τ between group information and each dataset increased, the correlation coefficient between multiset PLS-ROG score for response variable and each explanatory variable became higher. Consequently, we could confirm multiset PLS-ROG scores with rank order of groups clearly in this study of COVID-19 and hyperlipidemia. It was also commonly confirmed in COVID-19 and hyperlipidemia study that the multiset PLS-ROG scores swapped from the second multiset PLS-ROG score to the first multiset PLS-ROG score as the weight of connectivity between the groups and each dataset was increased. This also suggested that

the effect of rank order of groups might become clearer in first multiset PLS-ROG score as the weight of the connectivity between group and each dataset increased.

Finally, we would like to discuss how to set the parameter for the weight of connection between each omics datasets or response variable τ. In COVID-19 study, multiset PLS-ROG score for explanatory variables at τ=0.23 was affected by common outlier samples but could not show rank order of groups. On the other hands, multiset PLS-ROG score for explanatory variables at τ=0.25 was not affected by outliers and showed rank order of groups. When we focused on only rank order of groups, the optimal parameter of τ might be 0.25. However, in the case of human clinical research such as COVID-19 study, it may be practical to consider the rank order of groups with some allowance for outliers. In that case, the optimal parameter τ might be 0.24 in which both the effect of outliers and the order of the groups. As well as COVID-19 study, we confirmed the effects of change for the parameter τ for the weight of connection between each metabolome data on multiset PLS-ROG score in hyperlipidemia study. In addition, there may be also considered various patterns of connection between datasets in hyperlipidemia study, for example, in cases where the heart muscle was connected to the liver but not to the plasma. Thus, there are more pattens of connection between group and each dataset not only how the weight of connection is set, but also connected or not. In practical data analysis, it would be

important to subjectively determine the parameter τ for the weight of connection depending on the purpose. It will be necessary to further apply the multiset PLS-ROG to various data sets and further study how the parameter τ should be set.

5. Conclusion

We proposed multiset PLS-ROG to integrate multi-omics data with rank order of groups. We explored the effect of the parameters on the multiset PLS-ROG scores through practical data analysis of multi-omics data in COVID-19 study and multi-organ derived metabolome data in hyperlipidemia study. As a result, we found that both parameters of weight of connection τ between group and datasets and smoothing κ were important to extract scores with rank order of groups while only smoothing parameter is important in PLS-ROG. To identify compounds by statistical criteria, we clarified the statistical property of the weight vector in multiset PLS-ROG that was proportional to the weighted sum of correlation coefficient between multiset PLS-ROG score for response variable and each compound levels and multiset PLS-ROG score for explanatory variable and each compound levels. With this statistical property, we defined multiset PLS-ROG loading as the weighted correlation coefficient and could identify statistically significant compounds by using statistical hypothesis testing of multiset PLS-ROG loading.

CRediT author statement

Hiroyuki Yamamoto: Conceptualization, Methodology, Software, Writing original draft and reviewing and Editing.

Declaration of competing interest

The author is an employee of Human Metabolome Technologies, Inc.

References

[1] S. Huang, K. Chaudhary, L.X. Garmire, More is better: Recent progress in multi-omics data integration methods, Front. Genet. 8 (2017) 1–12. https://doi.org/10.3389/fgene.2017.00084.

[2] F.R. Pinu, D.J. Beale, A.M. Paten, K. Kouremenos, S. Swarup, H.J. Schirra, D. Wishart, Systems biology and multi-omics integration: Viewpoints from the metabolomics research community, Metabolites. 9 (2019) 1–31. https://doi.org/10.3390/metabo9040076.

[3] N. Ishii, K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M. Naba, K. Hirai, A. Hoque, P.Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T. Masuda, N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N. Iwata, Y. Toya, Y. Nakayama, T. Nishioka, K. Shimizu, H. Mori, M. Tomita, Multiple high-throughput analyses monitor the response of E. coli to perturbations, Science (80-.). 316 (2007) 593–597. https://doi.org/10.1126/science.1132067.

[4] K. Yugi, H. Kubota, A. Hatano, S. Kuroda, Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple “Omic” Layers, Trends Biotechnol. 34 (2016) 276–290. https://doi.org/10.1016/j.tibtech.2015.12.013.

[5] J. Lloyd-Price, C. Arze, A.N. Ananthakrishnan, M. Schirmer, J. Avila-Pacheco, T.W.

Poon, E. Andrews, N.J. Ajami, K.S. Bonham, C.J. Brislawn, D. Casero, H. Courtney, A. Gonzalez, T.G. Graeber, A.B. Hall, K. Lake, C.J. Landers, H. Mallick, D.R. Plichta, M. Prasad, G. Rahnavard, J. Sauk, D. Shungin, Y. Vázquez-Baeza, R.A. White, J. Bishai, K. Bullock, A. Deik, C. Dennis, J.L. Kaplan, H. Khalili, L.J. McIver, C.J. Moran, L. Nguyen, K.A. Pierce, R. Schwager, A. Sirota-Madi, B.W. Stevens, W. Tan, J.J. ten Hoeve, G. Weingart, R.G. Wilson, V. Yajnik, J. Braun, L.A. Denson, J.K. Jansson, R. Knight, S. Kugathasan, D.P.B. McGovern, J.F. Petrosino, T.S. Stappenbeck, H.S. Winter, C.B. Clish, E.A. Franzosa, H. Vlamakis, R.J. Xavier, C. Huttenhower, Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases, Nature. 569 (2019) 655–662. https://doi.org/10.1038/s41586-019-1237-9.

[6] Y. Hasin, M. Seldin, A. Lusis, Multi-omics approaches to disease, Genome Biol. 18 (2017) 1–15. https://doi.org/10.1186/s13059-017-1215-1.

[7] A. Csala, A.H. Zwinderman, "Multivariate Statistical Methods for High-Dimensional Multiset Omics Data Analysis", Computational Biology, Husi H, editor., Brisbane (AU), Codon Publications (2019). http://dx.doi.org/10.15586/computationalbiology.2019.ch5

[8] F. Rohart, B. Gautier, A. Singh, K.A. Lê Cao, mixOmics: An R package for 'omics feature selection and multiple data integration, PLoS Comput. Biol. 13 (2017) 1–19. https://doi.org/10.1371/journal.pcbi.1005752.

[9] H. Yamamoto, T. Fujimori, H. Sato, G. Ishikawa, K. Kami, Y. Ohashi, Statistical hypothesis testing of factor loading in principal component analysis and its application to metabolite set enrichment analysis, BMC Bioinformatics. 15 (2014). https://doi.org/10.1186/1471-2105-15-51.

[10] H. Yamamoto, PLS-ROG: Partial least squares with rank order of groups, J. Chemom. 31 (2017) e2883. https://doi.org/10.1002/cem.2883.

[11] H. Yamamoto, PLS-ROG: Partial least squares with rank order of groups, J. Chemom. 31 (2017) e2883. https://doi.org/10.1002/cem.2883.

[12] B. Shen, X. Yi, Y. Sun, X. Bi, J. Du, C. Zhang, S. Quan, F. Zhang, R. Sun, L. Qian, W. Ge, W. Liu, S. Liang, H. Chen, Y. Zhang, J. Li, J. Xu, Z. He, B. Chen, J. Wang, H. Yan, Y. Zheng, D. Wang, J. Zhu, Z. Kong, Z. Kang, X. Liang, X. Ding, G. Ruan, N. Xiang, X. Cai, H. Gao, L. Li, S. Li, Q. Xiao, T. Lu, Y. Zhu, H. Liu, H. Chen, T. Guo, Proteomic and Metabolomic Characterization of COVID-19 Patient Sera, Cell. 182 (2020) 59-72.e15. https://doi.org/10.1016/j.cell.2020.05.032.

[13] T. Ooga, H. Sato, A. Nagashima, K. Sasaki, M. Tomita, T. Soga, Y. Ohashi, Metabolomic anatomy of an animal model revealing homeostatic imbalances in dyslipidaemia, Mol. Biosyst. 7 (2011) 1217–1223. https://doi.org/10.1039/c0mb00141d.

[14] P. Amarenco, J. Bogousslavsky, A. 3rd Callahan, L.B. Goldstein, M. Hennerici, A.E. Rudolph, H. Sillesen, L. Simunovic, M. Szarek, K.M.A. Welch, J.A. Zivin, High-dose atorvastatin after stroke or transient ischemic attack., N. Engl. J. Med. 355 (2006) 549–559. https://doi.org/10.1056/NEJMoa061894.

[15] S.L. Rice, R.R. Eitenmiller, P.E. Koehler, Biologically Active Amines in Food: A Review, J. Milk Food Technol. 39 (1976) 353–358. https://doi.org/10.4315/0022-2747-39.5.353

[16] E. Sasaki, H. Yamamoto, T. Asari, R. Matsuta, S. Ota, Y. Kimura, S. Sasaki, K. Ishibashi, Y. Yamamoto, K. Kami, M. Ando, E. Tsuda, Y. Ishibashi, Metabolomics with severity of radiographic knee osteoarthritis and early phase synovitis in middle-aged women from the Iwaki Health Promotion Project: a cross-sectional study, Arthritis Res. Ther. 24 (2022) 1–11. https://doi.org/10.1186/s13075-022-02830-w.

Figure legend

Figure 1 Diagram of connection between dataset and group information. (A) Connections between each omics datasets, and between each omics dataset and group information (B) Connections between each omics dataset and group information only.

Figure 2 First and second multiset PLS-ROG scores in COVID-19 study. Three outlier samples of JBDZ2, JBDZ3 and JBDZ6 were blacked out. (A_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in proteome dataset. (○) Healthy control (HC), (△) mild COVID-19, (□) severe COVID-19. A box and whisker plot of (A_2) first multiset PLS-ROG score and (A_3) second multiset PLS-ROG score for explanatory variables to HC, mild and severe COVID-19 groups in proteome dataset. (B_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in metabolome dataset. A box and whisker plot of (B_2) first multiset PLS-ROG score and (B_3) second multiset PLS-ROG score for explanatory variables to HC, mild and severe COVID-19 groups in metabolome dataset. (C_1) A scatter plot of first and second multiset PLS-ROG score for response variables. A bar plot of (C_2) first multiset PLS-ROG score and (C_3) second multiset PLS-ROG score for response variable to HC, mild and severe COVID-19 groups.

Figure 3 Correlation coefficient between multiset PLS-ROG score for response and explanatory variables when $\tau_{1,Y}=\tau_{2,Y}$ are set from 0.3 to 0.5 in 0.01 increments. (○) Correlation coefficient between first multiset PLS-ROG score for explanatory variables in proteome and metabolome data,

(●) correlation coefficient between first multiset PLS-ROG score for response and explanatory variable in proteome data (■) and correlation coefficient between first multiset PLS-ROG score for response and explanatory variable in metabolome data.

Figure 4 A box and whisker plot of first multiset PLS-ROG score for explanatory variables to HC, mild and severe COVID-19 groups at (A) $\tau_{1,Y}=\tau_{2,Y}=0.48$, (B) 0.49 and (C) 0.50 in proteome dataset.

Figure 5 A box and whisker plot of top 3 ranked proteins, (A) P12111, (B) P13727, (C) P24821, (D) Q14520, (E) P12259 and (F) Q15113 by using multiset PLS-ROG loading at $\tau_{1,Y}=\tau_{2,Y}=0.48$, 0.49 and 0.50.

Figure 6 First and second multiset PLS-ROG scores ($\tau = 0.1$) in hyperlipidemia study. (○) Wild type (WT), (△) hyperlipidemia (HL) rabbit administered with drug, (□) HL rabbit. (A_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in plasma samples. A box and whisker plot of (A_2) first multiset PLS-ROG score and (A_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in plasma samples. (B_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in liver samples. A box and whisker plot of (B_2) first multiset PLS-ROG score (B_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in liver samples. (C_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in heart muscle samples. A box and whisker plot of (C_2) first multiset PLS-ROG score and

(C_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in heart muscle samples. (D_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in brain samples. A box and whisker plot of (D_2) first multiset PLS-ROG score and (D_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in brain samples. (E_1) A scatter plot of first and second multiset PLS-ROG score for response variables. A box and whisker plot of (E_2) first multiset PLS-ROG score and (E_3) second multiset PLS-ROG score for response variable to WT, HL administered with drug and HL.

Figure 7 Correlation coefficient between multiset PLS-ROG score for response and explanatory variables when τ is set from 0.1 to 0.25 in 0.01 increments. Mean of correlation coefficient between first multiset PLS-ROG score for explanatory variables in (○) plasma, (△) liver, (□) heart muscle, (◇) brain and other samples and correlation coefficient between first multiset PLS-ROG score for response and explanatory variable for (●) plasma, (▲) liver, (■) heart muscle and (◆) brain samples.

Figure 8 A box and whisker plot of first multiset PLS-ROG score for explanatory variables to wild type (WT), hyperlipidemia (HL) administered with drug and HL in (A_1) plasma samples (τ=0.23), (A_2) liver samples (τ=0.23), (A_3) heart muscle samples (τ=0.23), (A_4) brain samples (τ=0.23), in (B_1) plasma samples (τ=0.24), (B_2) liver samples (τ=0.24), (B_3) heart muscle samples (τ=0.24), (B_4) brain samples (τ=0.24), in (C_1) plasma samples (τ=0.25), (C_2) liver samples (τ=0.25), (C_3) heart muscle samples (τ=0.25) and (C_4) brain samples (τ=0.25). A bar plot of first multiset PLS-ROG score for

response variable in (A_5) τ=0.23, (B_5) τ=0.24 and (C_5) τ=0.25 to HC, mild and severe COVID-19 groups.

Figure 9 A box and whisker plot of (A) carnosine, (B) citrulline, (C) O-acetylserine, (D) N6,N6,N6-trimethyllysine, (E) N,N-dimethylglycine, (F) tyramine and (G) Gly-Gly to wild type (WT), hyperlipidemia (HL) administered with drug and HL in heart muscle samples.

Supplementary

Figure S1 First and second unsupervised multiset PLS scores in COVID-19 study. Three outlier samples of JBDZ2, JBDZ3 and JBDZ6 were blacked out. (A) A scatter plot of first and second unsupervised multiset PLS score for explanatory variables in proteome dataset. (○) Healthy control (HC), (△) mild COVID-19, (□) severe COVID-19. (B) A scatter plot of first and second unsupervised multiset PLS score for explanatory variables in metabolome dataset.

Figure S2 First and second multiset PLS scores ($\tau_{1,2}=\tau_{1,Y}=\tau_{2,Y}=1/3$) in COVID-19 study. Three outlier samples of JBDZ2, JBDZ3 and JBDZ6 were blacked out. (A_1) A scatter plot of first and second multiset PLS score for explanatory variables in proteome dataset. (○) Healthy control (HC), (△) mild COVID-19, (□) severe COVID-19. A box and whisker plot of (A_2) first multiset PLS score and (A_3) second multiset PLS score for explanatory variables to HC, mild and severe COVID-19 groups in proteome dataset. (B_1) A scatter plot of first and second multiset PLS score for explanatory variables in metabolome dataset. A box and whisker plot of (B_2) first multiset PLS score and (B_3) second multiset PLS score for explanatory variables to HC, mild and severe COVID-19 groups in metabolome dataset. (C_1) A scatter plot of first and second multiset PLS score for response variables. A bar plot of (C_2) first multiset PLS score and (C_3) second multiset PLS score for response variable to HC, mild and severe COVID-19 groups.

Figure S3 First and second multiset PLS-ROG scores ($\tau_{1,Y}=\tau_{2,Y}=0.48$) in COVID-19 study. Three

outlier samples of JBDZ2, JBDZ3 and JBDZ6 were blacked out. (A_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in proteome dataset. (○) Healthy control (HC), (△) mild COVID-19, (□) severe COVID-19. A box and whisker plot of (A_2) first multiset PLS-ROG score and (A_3) second multiset PLS-ROG score for explanatory variables to HC, mild and severe COVID-19 groups in proteome dataset. (B_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in metabolome dataset. A box and whisker plot of (B_2) first multiset PLS-ROG score and (B_3) second multiset PLS-ROG score for explanatory variables to HC, mild and severe COVID-19 groups in metabolome dataset. (C_1) A scatter plot of first and second multiset PLS-ROG score for response variables. A bar plot of (C_2) first multiset PLS-ROG score and (C_3) second multiset PLS-ROG score for response variable to HC, mild and severe COVID-19 groups.

Figure S4 First and second multiset PLS-ROG scores ($\tau_{1,Y}=\tau_{2,Y}=0.49$) in COVID-19 study. Three outlier samples of JBDZ2, JBDZ3 and JBDZ6 were blacked out. (A_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in proteome dataset. (○) Healthy control (HC), (△) mild COVID-19, (□) severe COVID-19. A box and whisker plot of (A_2) first multiset PLS-ROG score and (A_3) second multiset PLS-ROG score for explanatory variables to HC, mild and severe COVID-19 groups in proteome dataset. (B_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in metabolome dataset. A box and whisker plot of (B_2) first multiset PLS-ROG score and (B_3) second multiset PLS-ROG score for explanatory variables to HC, mild and severe

COVID-19 groups in metabolome dataset. (C_1) A scatter plot of first and second multiset PLS-ROG score for response variables. A bar plot of (C_2) first multiset PLS-ROG score and (C_3) second multiset PLS-ROG score for response variable to HC, mild and severe COVID-19 groups.

Figure S5 First and second multiset PLS-ROG scores ($\tau_{1,Y}=\tau_{2,Y}=0.50$) in COVID-19 study. Three outlier samples of JBDZ2, JBDZ3 and JBDZ6 were blacked out. (A_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in proteome dataset. (○) Healthy control (HC), (△) mild COVID-19, (□) severe COVID-19. A box and whisker plot of (A_2) first multiset PLS-ROG score and (A_3) second multiset PLS-ROG score for explanatory variables to HC, mild and severe COVID-19 groups in proteome dataset. (B_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in metabolome dataset. A box and whisker plot of (B_2) first multiset PLS-ROG score and (B_3) second multiset PLS-ROG score for explanatory variables to HC, mild and severe COVID-19 groups in metabolome dataset. (C_1) A scatter plot of first and second multiset PLS-ROG score for response variables. A bar plot of (C_2) first multiset PLS-ROG score and (C_3) second multiset PLS-ROG score for response variable to HC, mild and severe COVID-19 groups.

Figure S6 A box and whisker plot of 5-methyluridine to HC, mild and severe COVID-19 groups.

Figure S7 First and second multiset PLS-ROG scores ($\tau = 0.21$) in hyperlipidemia study. (○) Wild type (WT), (△) hyperlipidemia (HL) rabbit administered with drug, (□) HL rabbit. (A_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in plasma samples. A box and

whisker plot of (A_2) first multiset PLS-ROG score and (A_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in plasma samples. (B_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in liver samples. A box and whisker plot of (B_2) first multiset PLS-ROG score (B_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in liver samples. (C_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in heart muscle samples. A box and whisker plot of (C_2) first multiset PLS-ROG score and (C_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in heart muscle samples. (D_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in brain samples. A box and whisker plot of (D_2) first multiset PLS-ROG score and (D_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in brain samples. (E_1) A scatter plot of first and second multiset PLS-ROG score for response variables. A bar plot of (E_2) first multiset PLS-ROG score and (E_3) second multiset PLS-ROG score for response variable to WT, HL administered with drug and HL.

Figure S8 First and second multiset PLS-ROG scores ($\tau = 0.22$) in hyperlipidemia study. (○) Wild type (WT), (△) hyperlipidemia (HL) rabbit administered with drug, (□) HL rabbit. (A_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in plasma samples. A box and whisker plot of (A_2) first multiset PLS-ROG score and (A_3) second multiset PLS-ROG score for

explanatory variables to WT, HL administered with drug and HL in plasma samples. (B_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in liver samples. A box and whisker plot of (B_2) first multiset PLS-ROG score (B_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in liver samples. (C_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in heart muscle samples. A box and whisker plot of (C_2) first multiset PLS-ROG score and (C_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in heart muscle samples. (D_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in brain samples. A box and whisker plot of (D_2) first multiset PLS-ROG score and (D_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in brain samples. (E_1) A scatter plot of first and second multiset PLS-ROG score for response variables. A bar plot of (E_2) first multiset PLS-ROG score and (E_3) second multiset PLS-ROG score for response variable to WT, HL administered with drug and HL.

Figure S9 First and second multiset PLS-ROG scores ($\tau = 0.23$) in hyperlipidemia study. (○) Wild type (WT), (△) hyperlipidemia (HL) rabbit administered with drug, (□) HL rabbit. (A_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in plasma samples. A box and whisker plot of (A_2) first multiset PLS-ROG score and (A_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in plasma samples. (B_1) A scatter

plot of first and second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in liver samples. A box and whisker plot of (B_2) first multiset PLS-ROG score (B_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in liver samples. (C_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in heart muscle samples. A box and whisker plot of (C_2) first multiset PLS-ROG score and (C_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in heart muscle samples. (D_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in brain samples. A box and whisker plot of (D_2) first multiset PLS-ROG score and (D_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in brain samples. (E_1) A scatter plot of first and second multiset PLS-ROG score for response variables. A bar plot of (E_2) first multiset PLS-ROG score and (E_3) second multiset PLS-ROG score for response variable to WT, HL administered with drug and HL.

Figure S10 First and second multiset PLS-ROG scores ($\tau = 0.24$) in hyperlipidemia study. (○) Wild type (WT), (△) hyperlipidemia (HL) rabbit administered with drug, (□) HL rabbit. (A_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in plasma samples. A box and whisker plot of (A_2) first multiset PLS-ROG score and (A_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in plasma samples. (B_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables to WT, HL administered

with drug and HL in liver samples. A box and whisker plot of (B_2) first multiset PLS-ROG score (B_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in liver samples. (C_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in heart muscle samples. A box and whisker plot of (C_2) first multiset PLS-ROG score and (C_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in heart muscle samples. (D_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in brain samples. A box and whisker plot of (D_2) first multiset PLS-ROG score and (D_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in brain samples. (E_1) A scatter plot of first and second multiset PLS-ROG score for response variables. A bar plot of (E_2) first multiset PLS-ROG score and (E_3) second multiset PLS-ROG score for response variable to WT, HL administered with drug and HL.

Figure S11 First and second multiset PLS-ROG scores ($\tau = 0.25$) in hyperlipidemia study. (○) Wild type (WT), (△) hyperlipidemia (HL) rabbit administered with drug, (□) HL rabbit. (A_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in plasma samples. A box and whisker plot of (A_2) first multiset PLS-ROG score and (A_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in plasma samples. (B_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in liver samples. A box and whisker plot of (B_2) first multiset PLS-ROG score (B_3)

second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in liver samples. (C_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in heart muscle samples. A box and whisker plot of (C_2) first multiset PLS-ROG score and (C_3) second multiset PLS-ROG score of explanatory variables to WT, HL administered with drug and HL in heart muscle samples. (D_1) A scatter plot of first and second multiset PLS-ROG score for explanatory variables in brain samples. A box and whisker plot of (D_2) first multiset PLS-ROG score and (D_3) second multiset PLS-ROG score for explanatory variables to WT, HL administered with drug and HL in brain samples. (E_1) A scatter plot of first and second multiset PLS-ROG score for response variables. A bar plot of (E_2) first multiset PLS-ROG score and (E_3) second multiset PLS-ROG score for response variable to WT, HL administered with drug and HL.

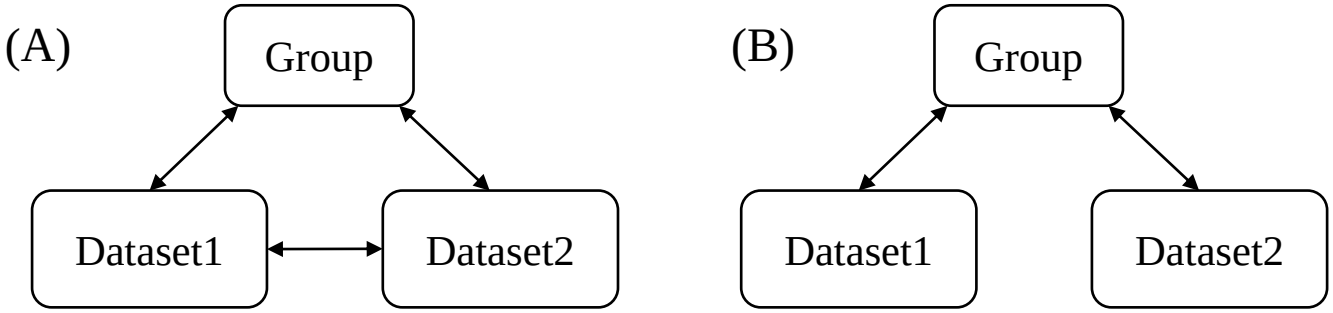

FIG.1

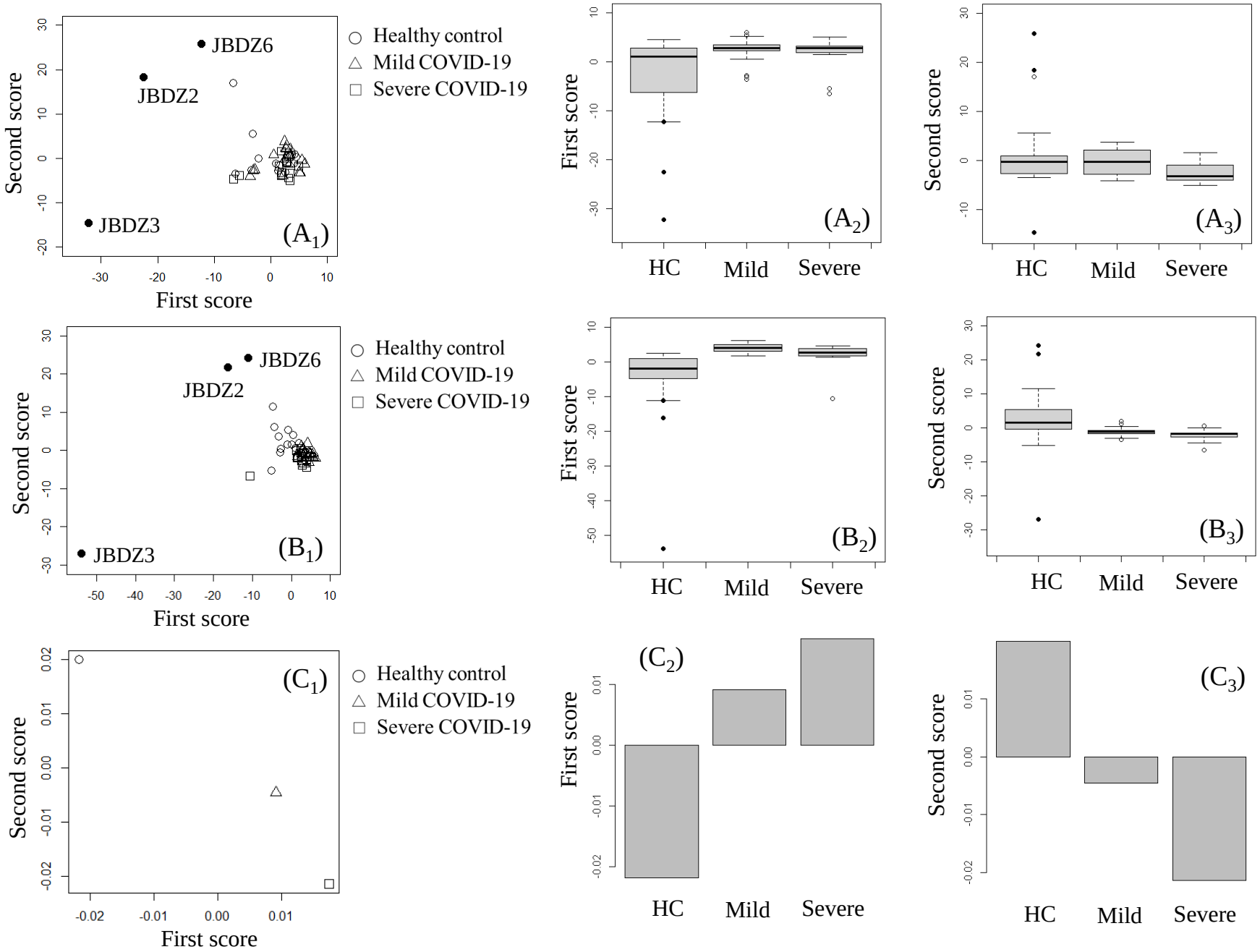
JBDZ6
JBDZ2
JBDZ3
(A1)
Healthy control
Mild COVID-19
Severe COVID-19
Second score
First score
(A2)
(A3)
HC
Mild
Severe
(B1)
(B2)
(B3)
(C1)
(C2)
(C3)

FIG.2

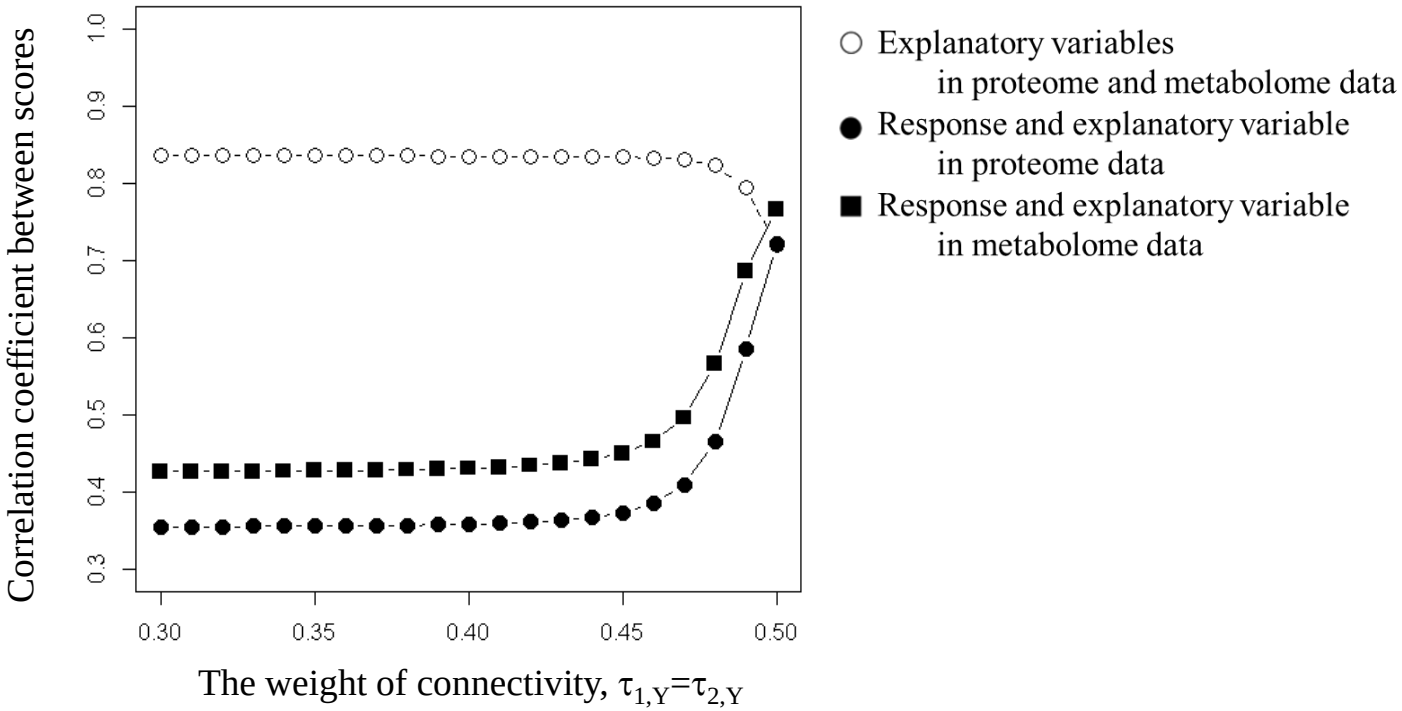
Correlation coefficient between scores
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.30
0.35
0.40
0.45
0.50
The weight of connectivity, $\tau_{1,Y}=\tau_{2,Y}$
Explanatory variables in proteome and metabolome data
Response and explanatory variable in proteome data
Response and explanatory variable in metabolome data

FIG.3

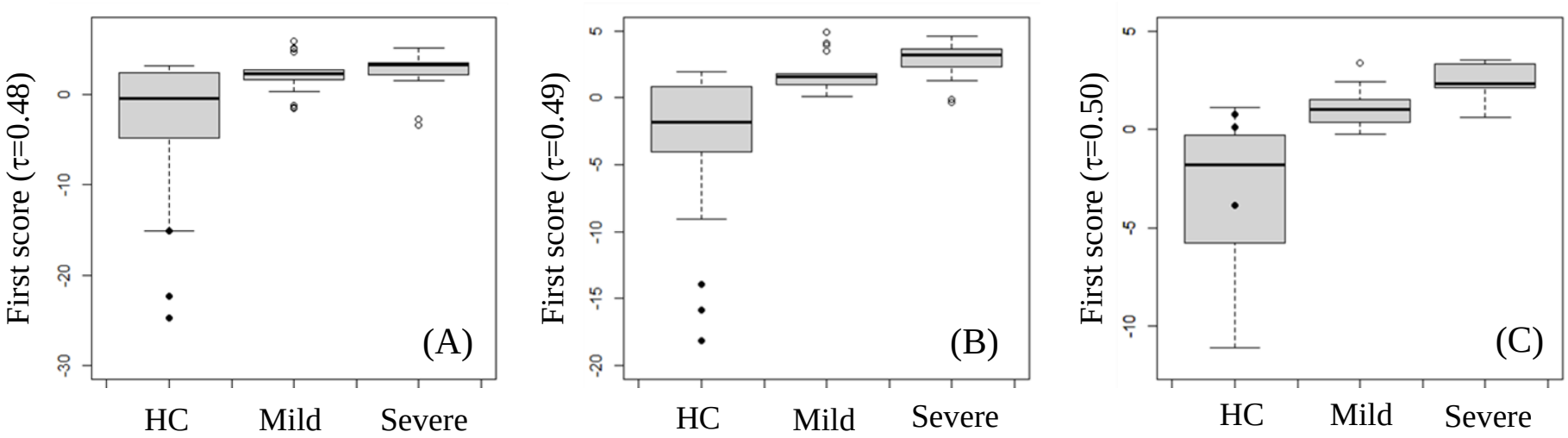
First score (τ=0.48)
0
-10
-20
-30
(A)
HC
Mild
Severe
First score (τ=0.49)
5
0
-5
-10
-15
-20
(B)
HC
Mild
Severe
First score (τ=0.50)
5
0
-5
-10
(C)
HC
Mild
Severe

FIG.4

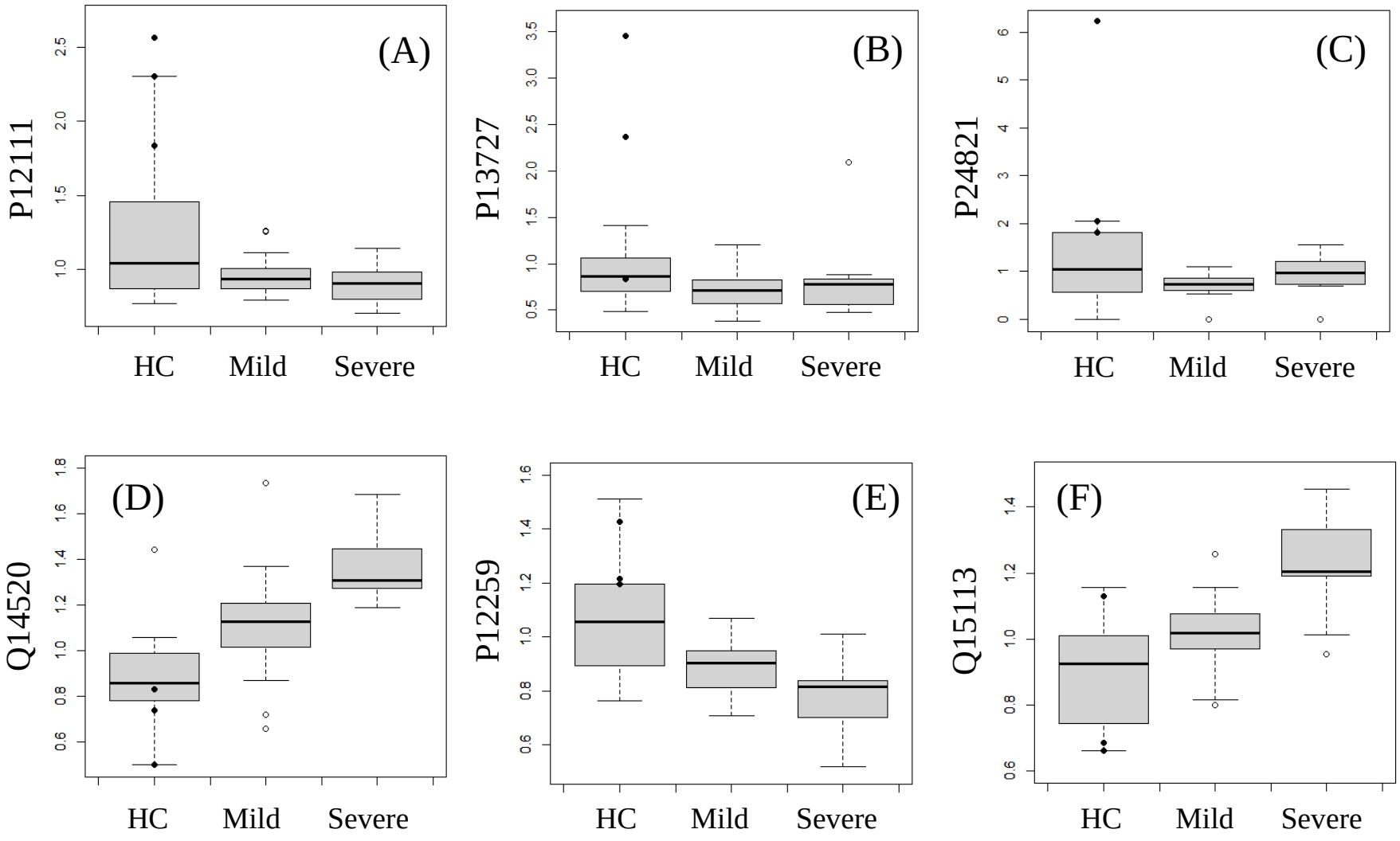
(A)
P12111
1.0
1.5
2.0
2.5
HC
Mild
Severe
(B)
P13727
0.5
1.0
1.5
2.0
2.5
3.0
3.5
HC
Mild
Severe
(C)
P24821
0
1
2
3
4
5
6
HC
Mild
Severe
(D)
Q14520
0.6
0.8
1.0
1.2
1.4
1.6
1.8
HC
Mild
Severe
(E)
P12259
0.6
0.8
1.0
1.2
1.4
1.6
HC
Mild
Severe
(F)
Q15113
0.6
0.8
1.0
1.2
1.4
HC
Mild
Severe

FIG.5

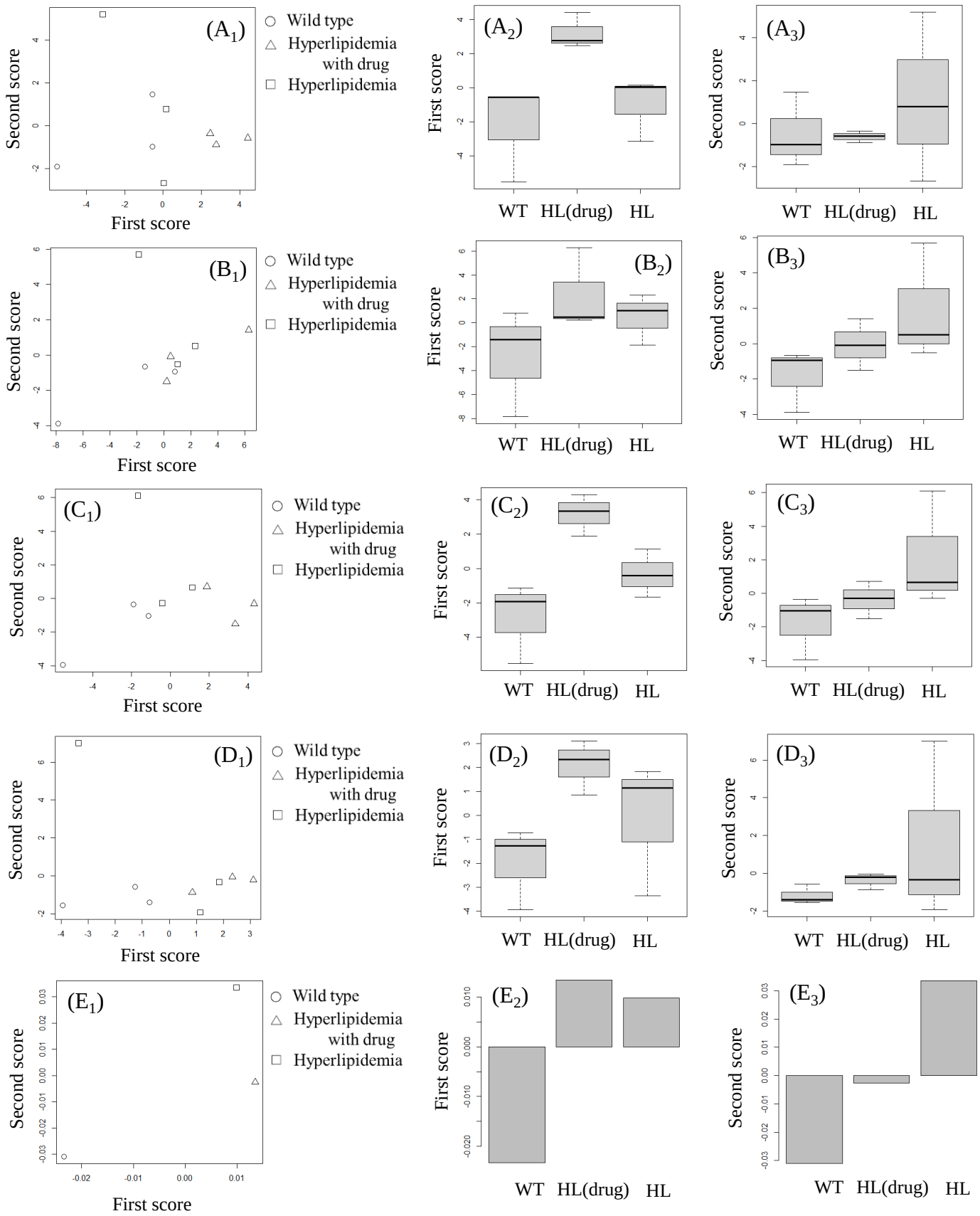
(A1)
(A2)
(A3)
(B1)
(B2)
(B3)
(C1)
(C2)
(C3)
(D1)
(D2)
(D3)
(E1)
(E2)
(E3)
Wild type
Hyperlipidemia with drug
Hyperlipidemia
First score
Second score
WT
HL(drug)
HL

FIG.6

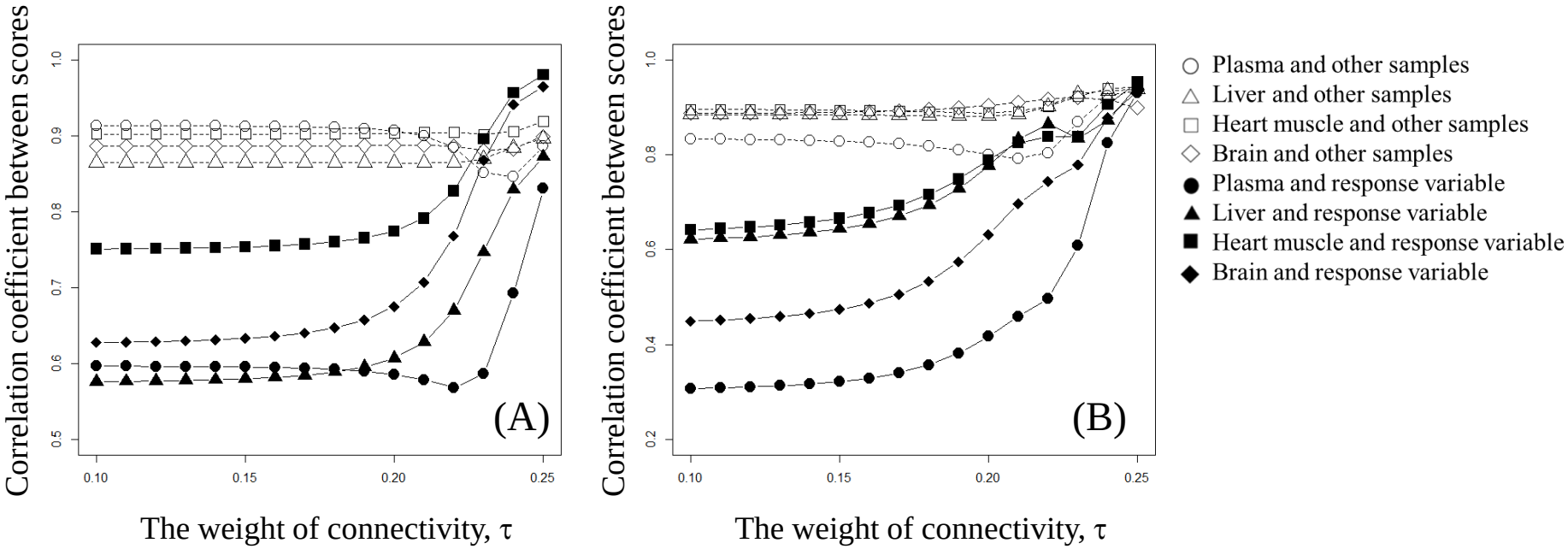
Correlation coefficient between scores
1.0
0.9
0.8
0.7
0.6
0.5
0.10
0.15
0.20
0.25
(A)
The weight of connectivity, τ
Correlation coefficient between scores
1.0
0.8
0.6
0.4
0.2
0.10
0.15
0.20
0.25
(B)
The weight of connectivity, τ
Plasma and other samples
Liver and other samples
Heart muscle and other samples
Brain and other samples
Plasma and response variable
Liver and response variable
Heart muscle and response variable
Brain and response variable

FIG.7

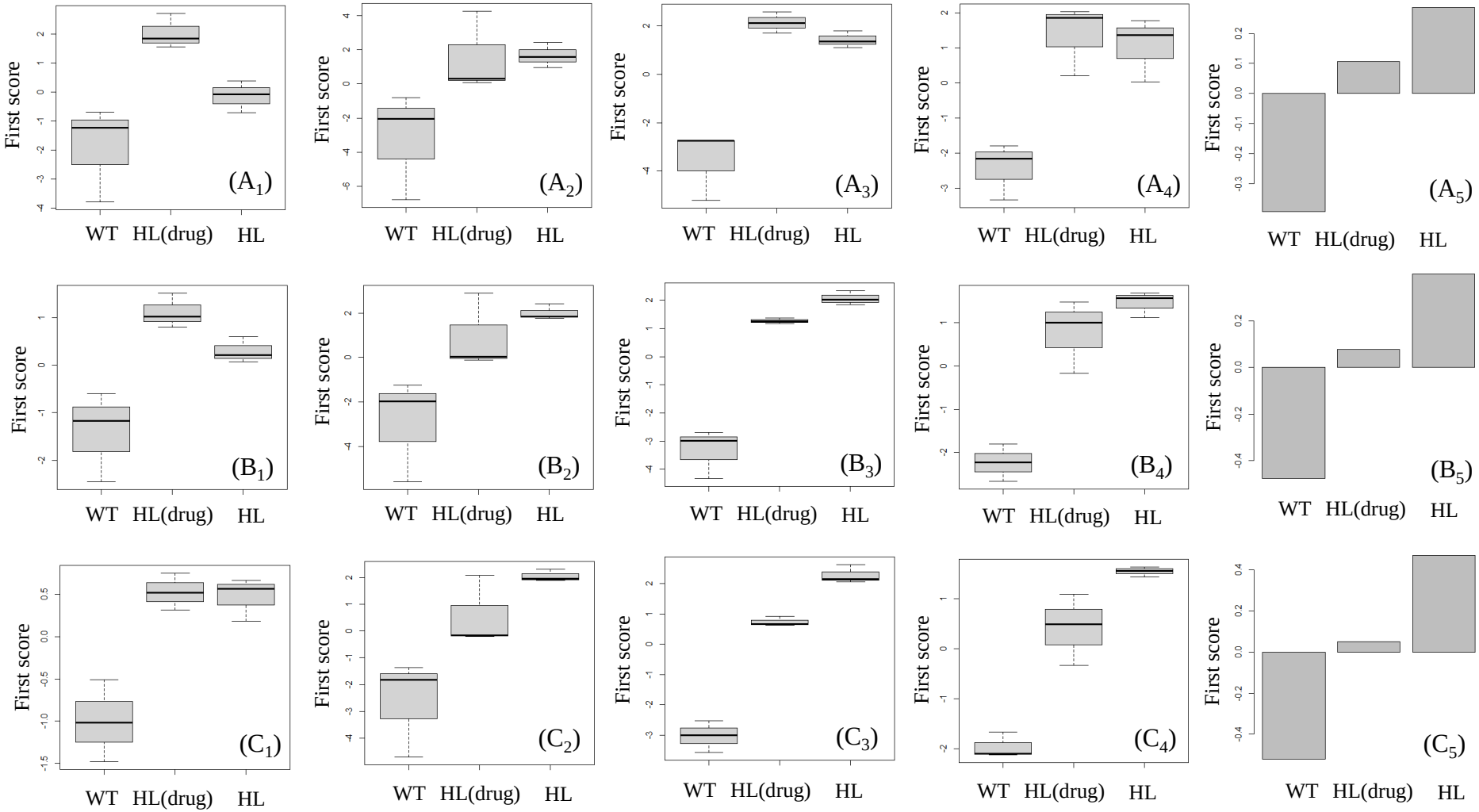
First score
WT
HL(drug)
HL
(A1)
(A2)
(A3)
(A4)
(A5)
(B1)
(B2)
(B3)
(B4)
(B5)
(C1)
(C2)
(C3)
(C4)
(C5)

FIG.8

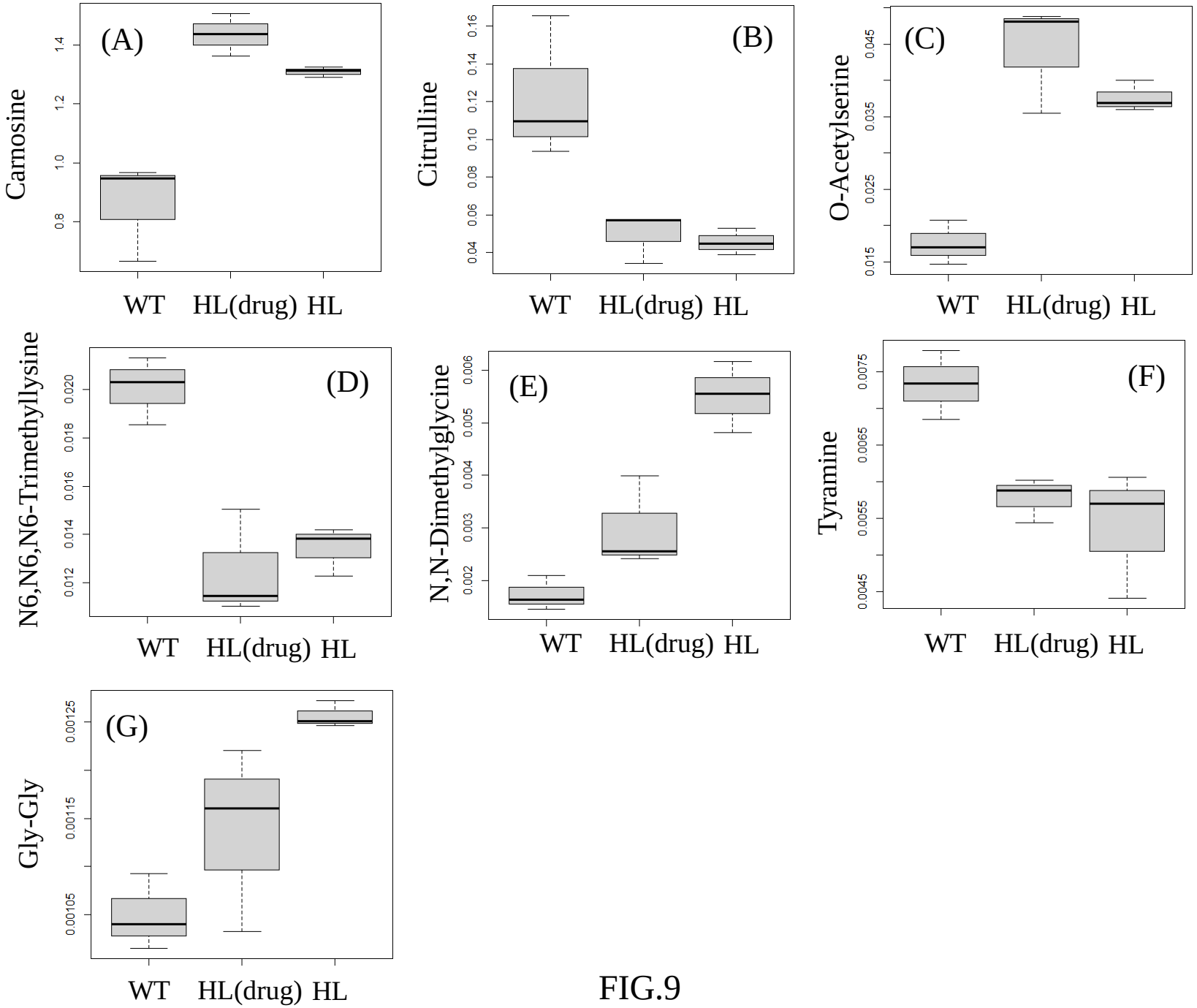
(A)
Carnosine
0.8
1.0
1.2
1.4
WT
HL(drug)
HL
(B)
Citrulline
0.04
0.06
0.08
0.10
0.12
0.14
0.16
WT
HL(drug)
HL
(C)
O-Acetylserine
0.015
0.025
0.035
0.045
WT
HL(drug)
HL
(D)
N6,N6,N6-Trimethyllysine
0.012
0.014
0.016
0.018
0.020
WT
HL(drug)
HL
(E)
N,N-Dimethylglycine
0.002
0.003
0.004
0.005
0.006
WT
HL(drug)
HL
(F)
Tyramine
0.0045
0.0055
0.0065
0.0075
WT
HL(drug)
HL
(G)
Gly-Gly
0.00105
0.00115
0.00125
WT
HL(drug)
HL

FIG.9