analytical chemistry

Article

pubs.acs.org/ac

Building Multivariate Systems Biology Models

Gemma M. Kirwan,*[,†] Erik Johansson,[‡] Robert Kleemann,[§] Elwin R. Verheij,[§] Åsa M. Wheelock,[∥] Susumu Goto,[†] Johan Trygg,[⊥] and Craig E. Wheelock*[,†,¶]

[†]Bioinformatics Centre, Institute for Chemical Research, Kyoto University, Kyoto, Japan

[‡]MKS Umetrics, Umeå, Sweden

[§]Metabolic Health Research, TNO, Zernikedreef 9, Leiden, The Netherlands

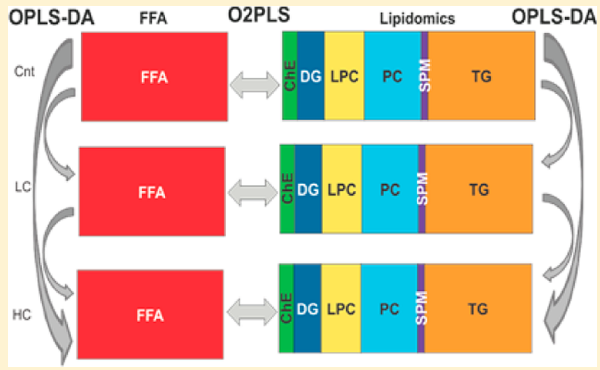
[∥]Respiratory Medicine Unit, Department of Medicine, Karolinska Institutet, Stockholm, Sweden

[⊥]Computational Life Science Cluster, Department of Chemistry, Umeå University, Umeå, Sweden

[¶]Department of Medical Biochemistry and Biophysics, Division of Physiological Chemistry II, Karolinska Institutet, Stockholm, Sweden

Supporting Information

ABSTRACT: Systems biology methods using large-scale "omics" data sets face unique challenges: integrating and analyzing near limitless data space, while recognizing and removing systematic variation or noise. Herein we propose a complementary multivariate analysis workflow to both integrate "omics" data from disparate sources and analyze the results for specific and unique sample correlations. This workflow combines principal component analysis (PCA), orthogonal projections to latent structures discriminate analysis (OPLS-DA), orthogonal 2 projections to latent structures (O2PLS), and shared and unique structures (SUS) plots. The workflow is demonstrated using data from a study in which ApoE3Leiden mice were fed an atherogenic diet consisting of increasing cholesterol levels followed by therapeutic intervention (fenofibrate, rosuvastatin, and LXR activator T-0901317). The levels of structural lipids (lipidomics) and free fatty acids in liver were quantified via liquid chromatography–mass spectrometry (LC–MS). The complementary workflow identified diglycerides as key hepatic metabolites affected by dietary cholesterol and drug intervention. Modeling of the three therapeutics for mice fed a high-cholesterol diet further highlighted diglycerides as metabolites of interest in atherogenesis, suggesting a role in eliciting chronic liver inflammation. In particular, O2PLS-based SUS2 plots showed that treatment with T-0901317 or rosuvastatin returned the diglyceride profile in high-cholesterol-fed mice to that of control animals.



The advent of high-throughput analytical technology platforms and the so-called "omics" revolution has resulted in the ability to produce biological data sets of enormous size and complexity.[1] These large data sets, coupled with complex biological questions, have driven the need for statistical, mathematical, and high-end computational support. Systems biology refers to a shift in the data analysis paradigm from traditional reductionism approaches to focus on the simultaneous study of the interrelationships of all elements in a system.[2] Current systems biology approaches face the challenge of establishing novel ways to integrate and analyze large data sets derived from multiple "omics" platforms. Commonly, the structure of these data sets consists of relatively few samples, but a large number of variable measurements, resulting in unique statistical challenges.

One of the primary challenges within systems biology remains the analysis of the large dimensional data space of high-throughput "omics" technologies. This task is made more difficult with the knowledge that most biological data sets contain both technical and biological noise and/or various types of systematic error. In the case of data space, not all measurements are necessarily relevant to the experimental hypothesis or system being explored and can potentially be excluded. Accordingly, the challenge consists of creating systems biology workflows that can offer data integration, identification of errors, and finally analysis with regard to hypothesis testing. To address these issues, a number of methodologies have emerged that are based on both new and classical multivariate analysis techniques. One in particular, orthogonal 2 projections to latent structures (O2PLS; an extension of partial least-squares), has the potential to integrate and analyze "omics" data sets while separating systematic

Received: May 15, 2012
Accepted: July 20, 2012
Published: July 20, 2012

   dx.doi.org/10.1021/ac301269r | *Anal. Chem.* 2012, 84, 7064–7071

variations and noise as well as unique structures unrelated to the other blocks from the joint correlation.

Partial least-squares/projections to latent structures (PLS) is commonly employed to examine variables for underlying correlations and patterns.[3] PLS can be considered unidirectional in that it only models *Y* from *X*. Orthogonal projections to latent structures (OPLS) is a PLS extension in which the systemic variation that is not correlated between the predictive and observed variables is modeled separately.[4] Although both OPLS and PLS have the same theoretical predictive power, the benefit of OPLS lies in its interpretability,[5] aiding in the selection of outliers, determining the number of components, and removing systematic variation from the predictive component. OPLS can also be employed as a discriminate analysis (OPLS-DA), describing differences between overall class properties while removing systematic variation.[6] The bidirectional correlations that exist in large multiblock data sets (the combination of "blocks" of data that represent subdomains or subgroups within a data set) make PLS methods unsuitable to describe multiblock correlations, and traditionally other bidirectional techniques such as canonical correlation (CC) and multiple correspondence analysis (MCA) have been employed, although they are considered to lack the proper model structure.[7] O2PLS is a new method of data integration and analysis that supports multiblock bidirectional correlations. O2PLS is an extension of OPLS and is both an exploratory and predictive bidirectional modeling tool. Generally, O2PLS produces three main outputs, the joint correlation that exists between *X* and *Y*, the specific variation in *X*, and the specific variation in *Y*.[8] O2PLS can also extend beyond two data sets and can be used to integrate and analyze large multiblock data sets of different units or sources.[9] O2PLS models assume joint variation or dependency in that any system under study must contain both independent and dependent variables (i.e., mRNA translation into a protein and/or a given protein effecting a shift in metabolite concentrations). To aid in the interpretation of OPLS models a visualization tool known as the SUS plot (shared and unique structures) plot can be employed.[10] We propose an extension of this tool for aiding in the interpretation of O2PLS models, herein referred to as an SUS2 plot.

To date there are relatively few literature examples of studies using O2PLS; however, it has been used to integrate and model plant transcript and metabolite data[11,12] and also been proposed as a viable statistical method for linking either genomic, transcriptomic, and/or proteomic variation with metabolomic variation.[13] By contrast OPLS (and PCA) related literature are prolific,[14−21] including application as a data dimensionally reduction tool.[22] In addition, OPLS-DA has also been gaining momentum.[23−25]

This paper details a novel, systematic approach to multivariate analysis for large-scale, multiblock studies. Multiblock data typifies the case where a number of variables have been measured across different samples, and the subsets of these measurements (gene profiles, metabolites, etc.) constitute different data blocks.[26] A demonstration of our proposed workflow is given using a data set containing measurements from two analytical platforms (lipidomics and free fatty acids) from ApoE3Leiden mice fed an atherogenic diet.[27,28] The workflow is first applied to examine the effects of low and high atherogenic dietary regimes upon hepatic metabolites and then extended to analyze the metabolic effects of three atherogenic therapeutic interventions, the PPAR-α activator fenofibrate (FF), the HMG CoA-reductase inhibitor rosuvastatin (RSV), and the LXR-activator T-0901317 (T09).

METHOD

Data Analysis. All data processing, including PCA, OPLS-DA, O2PLS, and cross validation, was implemented using SIMCA software v.13 (Umetrics AB, Umeå, Sweden).[8,11] Data were column centered variance scaled prior to multivariate analysis. O2PLS is an extension of OPLS that is not yet widely employed and is therefore briefly described here. Whereas OPLS is a unidirectional modeling technique (*Y* is predicted from *X*), O2PLS is bidirectional; *Y* can predict *X* and *X* can also predict *Y*, and uncorrelated or specific variation in both the *X* and *Y* blocks is removed as orthogonal components. This means that an O2PLS analysis produces multiple model components including the joint variation that exists between *X* and *Y* and the individual or specific orthogonal variation that exists purely in *X* and again in *Y* (Figure 1). This paper

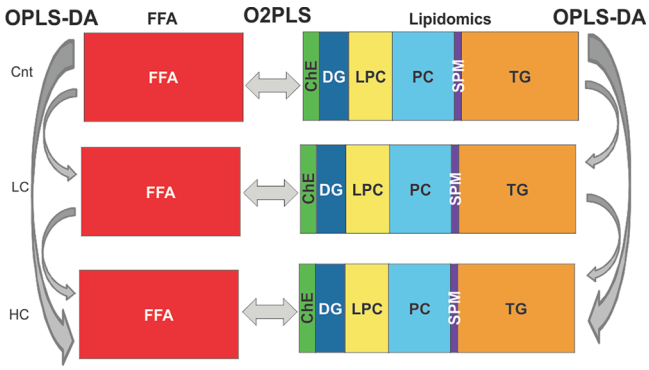


Figure 1. Overview of data models for dietary regimes (control [Cnt], low cholesterol [LC], and high cholesterol [HC]); "unfolding the data". ChE is cholesterol ester(s), DG is diglyceride(s), TG is triglyceride(s), LPC is lysophosphatidylcholine(s), PC is phosphatidylcholine(s), SPM is sphingomyelin, and FFA is free fatty acid(s).

demonstrates a single-class approach where *X* and *Y* only contain samples from one class of data (e.g., control). However, it is also possible to utilize a two-class approach, wherein *X* and *Y* contain two classes (e.g., control and experimental intervention). Two-class O2PLS has been previously applied to biological studies.[11,29] The user should consider which approach is most appropriate/insightful for a given study design. The current approach using single groups in the O2PLS analysis is primarily useful for cases where subgroups are suspected as in the presented data set. Values of R^2 and Q^2 are commonly quoted as descriptors of a multivariate model. R^2 represents the percentage of variation within a data set that can be explained by the model, often referred to as a measure of fit. The Q^2 is percent variation of the response predicted by the model according to cross validation, or rather, how accurately the model can be expected to predict new data. Since O2PLS is a bidirectional model between an *X* and *Y* block, SIMCA 13 calculates R^2X and R^2Y as the fraction of the sum of squares of all *X* and *Y* that the model can explain using the latent variables. These values for all models are shown in Supporting Information Table S1.

Samples. A total of 37 12 week old female ApoE3Leiden mice were separated into three dietary regulated groups: control, low cholesterol (LC), and high cholesterol (HC). The control group (n = 13) was fed a cholesterol-free diet for 10

dx.doi.org/10.1021/ac301269r | Anal. Chem. 2012, 84, 7064−7071

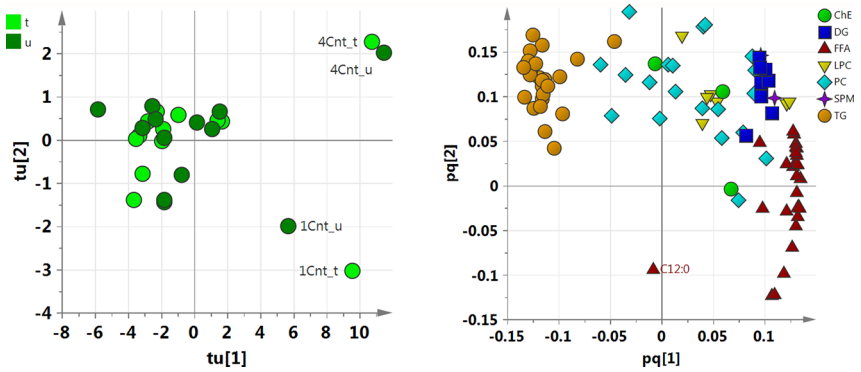


Figure 2. O2PLS joint scores plot (left) of control (Cnt), where *t* represents the scores from the *X* block and *u* represents the scores from the *Y* block, and O2PLS joint loadings plot (right) of control (Cnt), where *p* represents the loadings from the *X* block and *q* represents the loadings from the *Y* block; C12:0 is lauric acid. Lipid species nomenclature is as described in the text.

weeks. The low-cholesterol (n = 12) and high-cholesterol groups (n = 12) underwent the same 10 week dietary regime as the control, but their diets were supplemented with an additional 0.25% and 1% w/w cholesterol, respectively. After the 10 week feeding regime, mice were euthanized collectively under anesthesia and livers immediately collected, snap-frozen in liquid nitrogen, and stored at −80 °C until required for free fatty acid and lipidomics analysis.

A separate cohort of 36 12 week old female ApoE3Leiden mice undergoing therapeutic intervention were separated into three groups (n = 12 each), fed the same high-cholesterol diet as the HC group, but supplemented with one of the following: PPAR-α activator fenofibrate (n = 12, FF; 0.03% w/w), HMG CoA-reductase inhibitor rosuvastatin (n = 12, RSV; 0.05% w/w), or LXR-activator T-0901317 (n = 12, T09; 0.01% w/w). After the 10 week feeding/therapeutic intervention regime, mice were euthanized collectively under anesthesia and livers immediately collected, snap-frozen in liquid nitrogen, and stored at −80 °C until required for free fatty acid and lipidomics analysis.

Liver Free Fatty Acid and Lipidomics Analysis. Lipid and free fatty acid extraction: livers were homogenized, and 5 μL of homogenate was extracted with 200 μL of isopropyl alcohol containing internal standards (heptadecanoyl-lysophosphatidylcholine, dilauroyl-phosphatidylcholine, heptadecanoyl-cholesterol, and triheptadecanoyl-glycerol, all at a concentration of 1 μg/mL; Sigma, St. Louis, MO, U.S.A.) as previously described.[27]

Data were acquired by electrospray liquid chromatography−mass spectrometry (ESI-LC−MS) as described in the Supporting Information. Lipidomics data included these general lipid classes: cholesterol esters (ChE), diglycerides (DG), triglycerides (TG), lysophosphatidylcholines (LPC), phosphatidylcholines (PC), and sphingomyelin (SPM). Free fatty acid data (FFA) included 28 fatty acids ranging from C:12 to C:24 carbon chain length. The lipidomics data for the cholesterol dietary feeding study were previously published,[27] but the free fatty acid data and therapeutic drug intervention studies consist of novel data.

RESULTS AND DISCUSSION

This section details a systematic approach for a multivariate workflow through the complementary integration and analysis of data using PCA, OPLS-DA, and O2PLS (see Supporting Information Figure S1 for a workflow diagram). For simplicity, the proposed workflow initially focuses on the dietary effects of an atherogenic diet (cholesterol-free control [Cnt], low cholesterol [LC], high cholesterol [HC]). Later the study is extended to the analysis of three antiatherogenic therapeutic interventions. The proposed workflow involves the generation of multiple intermediate models during the analysis. This information is useful for understanding the full procedure, but is beyond the scope of the current paper. Accordingly, this material related to the various OPLS-DA and O2PLS models is provided for interested readers in the Supporting Information as Figures S2−S11, but is not discussed in detail in the main text.

The analysis workflow involves multiple steps; initially a data overview is performed (PCA and O2PLS), where data trends are examined and outliers or unique sample variation is observed (step 1). This step is followed by the "yardstick" (step 2), where specific variation within and between control samples is observed and a reference model is created for all other analyses. Exploratory and predictive analysis (step 3) describes the complementary use of O2PLS and OPLS-DA to explore the causality of the experiment. Figure 1 provides an overview of the O2PLS and OPLS-DA models created during exploratory and predictive analysis and the respective FFA and lipidomics data blocks used for each analysis. Finally the various OPLS-DA and O2PLS models are compared using SUS and SUS2 plots, respectively, to highlight significant differences or changes in samples and variables (step 4), which leads to the biological interpretation of results (step 5).

Step 1: Generate an Overview of the Data Set. The FFA and lipidomics data were divided into two blocks for O2PLS analysis. The decision on the block division was made on the basis of the following argument: fatty acids are esterified into varying head groups (i.e., triglycerides, phospholipids) to form different lipid classes, which can be released again via lipase activity. Accordingly, there is biological cross-talk between the two blocks given that FFAs are incorporated into structural lipids, which could be affected by the treatment with escalating doses of dietary cholesterol.

The initial overview analysis employed O2PLS and PCA on the complete FFA and lipidomics data sets (containing results from the Cnt, LC, and HC diets; Supporting Information Figure S2) to screen for unsupervised groupings and provide an overview of the data. There is a different presentation style of the O2PLS scores plot from traditional scores plots. As O2PLS is a bidirectional modeling technique, two equally important score vectors are created (**t** and **u**), and therefore, both are displayed in the scatter plot. Accordingly, there is double the

dx.doi.org/10.1021/ac301269r | Anal. Chem. 2012, 84, 7064−7071

number of samples; including both the **t** and **u** vectors. Of note is the fact that both models (**t** and **u**) arrive at and display the same data patterns and groupings, albeit with slightly different locations of sample data points. This is expected and attributed to the O2PLS modeling using blocks of data with differing origin variables. Figure 2 shows this style of scores plot (26 data points; 13 from the **t** vector and 13 from the **u** vector), and a tight cluster around the origin can be seen with two samples deviating from this mean cluster.

To empirically test the O2PLS *X* and *Y* blocks employed (FFA and lipidomics, respectively), the analysis was repeated using different combinations of *X* and *Y* data structures (e.g., FFA and TG vs lipidomics TG and DG vs FFA and lipidomics). The models produced from these alternative data structure experiments were not as strong (determined using cross validation), or no model could be obtained. The user should consider the design of the experiment[30] carefully, and the process of causality, before determining which data blocks to use.

Step 2: "Yardstick", Using PCA and O2PLS To Generate Baseline Models. Following the initial analysis, the yardstick analysis provides a reference for intersample biological variation and therefore a yardstick measurement to determine a significant shift (in this example, within the context of dietary regime and later therapeutic treatment-associated metabolic change). The yardstick analysis is performed using only control data (Figure 2 and Supporting Information Figure S3). The approach of single-group O2PLS is applicable only in specific cases where subgroups are present, or alternatively where one block is constant between groups. PCA was conducted on the control FFA and control lipidomics data (figures not shown), and O2PLS was used to examine the joint variation (Figure 2) between control FFA and lipidomics data. PCA is once again used as a precautionary analysis to determine if variables are grouping. The O2PLS joint variation revealed that two control samples (1Cnt, 4Cnt) displayed deviation from the central cluster.

The O2PLS control loadings plot (Figure 2) showed tight clustering groups of TGs, DGs, and FFAs, with some group overlap. Conversely, the ChE, LPC, PC, and SPM variables showed a large amount of intersample variability and did not form tight groupings, but rather spread out around the plot origin. The two outlying control samples (1Cnt, 4Cnt) appeared to have a significant difference in their PC and DG concentrations in comparison to the other controls. Examination of the specific variation of the control FFA and lipidomics data revealed that certain individual variables had a large amount of specific (intersample) variation, and the outlying sample (12Cnt) had significantly higher concentrations of TGs.

Whether or not these discrepancies are due to experimental error or represent biological variability requires further investigation. The ApoE3Leiden mouse carries a human transgene that encodes for two proteins, a mutated form of ApoE3 and ApoC1,[31] and expression of both proteins is relatively variable among littermates. Since both proteins can modulate plasma TG levels, it is most likely that the observed variance in TGs for sample 12Cnt is biologically related. The yardstick step revealed useful information regarding the variability of the control samples and generated the necessary control O2PLS models for further comparisons.

Step 3: Exploratory and Predictive Analysis. The bulk of the data analysis is performed in this step, employing a combination of exploratory O2PLS and predictive OPLS-DA to gain insight into biological changes due to different dietary regimes (refer to Figure 1 for all O2PLS and OPLS-DA combinations of experimental data).

O2PLS exploratory: control models were discussed in the yardstick analysis. The O2PLS LC FFA versus LC lipidomics scores plot showed tight joint variation, with most samples clustering around the origin. The loadings plot showed tighter grouping of PCs, SPMs, and FFAs than those observed in the control. The DG cluster spread out, becoming more decentralized, and the TGs shifted in their mean position. Two samples (7LC and 1LC) were located far from the central cluster, and examination of the loadings showed that they possessed distinct differences in their TG and PC profiles. The specific (orthogonal) components displayed a large amount of variation in one particular sample (9LC), which had large biological variation in TG concentrations over time. These observations can be seen in the Supporting Information (Figure S4).

The O2PLS HC FFA versus HC lipidomics results followed a similar trend, although the sample cluster in the scores plot was not as tight as in the LC and control analysis. The overlaid joint variation between HC FFA and HC lipidomics derived using predictive blocks *X* (FFA) versus *Y* (lipidomics) and *Y* versus *X*, respectively. Two samples (4HC, 7HC) were again noted as being distant from the central cluster.

The loadings plot of HC FFA versus HC lipidomics diet showed that (in comparison to the other diets) specific TGs had moved considerably from the mean grouping, and the DGs no longer grouped. Specific PCs had also left the main clusters, while ChEs began to group closely, indicating stabilization of mean concentrations. The loadings plot revealed that the two outlying samples (4HC, 7HC) were located right of the mean cluster due to FFAs, ChEs, LPCs, and SPMs. After inspecting the original data for these samples it was observed that they both shared abnormally high (with regard to the mean) concentrations of FFAs and ChEs. The O2PLS HC FFA versus HC lipidomics model contained only one specific component, which displayed large specific (unique) variation in sample 11HC attributed to its unique DG concentration profiles and reduced ChE concentrations (with regard to the group mean). These observations can be viewed in the Supporting Information (Figure S5).

OPLS-DA predictive: the OPLS-DA analysis predicts which class (group) a sample belongs to based on the information contained in its measured variables. The FFA data in itself could not produce OPLS-DA models; models were deemed to have too low R^2 and Q^2 values to be acceptable for prediction or analysis (based on cross-validation rules outlined in SIMCA 13).

OPLS-DA models of Cnt lipidomics versus LC lipidomics provided good predictability, with clear definition between the two groups. The loading plot showed that separation of LC samples occurred based primarily on PCs and ChEs. One sample was observed to lie outside of the mean cluster (4Cnt). OPLS-DA control lipidomics versus LC lipidomics orthogonal components showed two samples with unique variation (1LC, 7LC).

The OPLS-DA model of Cnt lipidomics versus HC lipidomics showed clear separation between groups. The same sample noted previously during the yardstick analysis (12Cnt) as being higher in TGs was now easily identified, again separating from the main control group based on high TG concentrations. The loadings plot showed that the main

dx.doi.org/10.1021/ac301269r | *Anal. Chem.* 2012, 84, 7064−7071

variables that separated the HC group from the control, the predictive variables, were again PCs and ChEs. The specific variation again highlighted the one high TG sample (12Cnt), as well as another sample (4Cnt) with systematic variation in PCs.

OPLS-DA models of LC versus HC lipidomics also demonstrated good predictability and clear separation between samples. Examination of the loadings plot showed that the driving factors separating LC and HC samples were PCs and ChEs. Sample 1LC was again observed to separate from the main LC cluster. Orthogonal components once more showed samples 7HC, 1LC, and 7LC as having unique variation, as well as sample 11HC. Sample 11HC shifted from the main HC cluster due to variance in DGs and PCs. Supporting Information contains the OPLS-DA plots (Figures S6–S9).

Step 4: Comparison of Models and Reconstructing the Data Using SUS and SUS2 Plots. The SUS and SUS2 plots compare the outcome of various models, relative to a reference (usually control), visualizing the variable influences between models. These scatter plots combine the correlation profiles from two multivariate models and display correlations scaled between −1 and 1 on both axes.[10]

Like all visual interpretations, the analysis of SUS (and SUS2) plots for the selection of interesting variables is subjective and requires prior knowledge of the experimental design. The SUS (and SUS2) plot displays variables of the same correlation (but from different models) along a diagonal. The further from the diagonal a variable lies the greater its influence between models and hence the dynamic changes occurring in the variable between the models.

The SUS2 plot comparing HC and Cnt models (Figure 3) showed pronounced changes between levels of hepatic

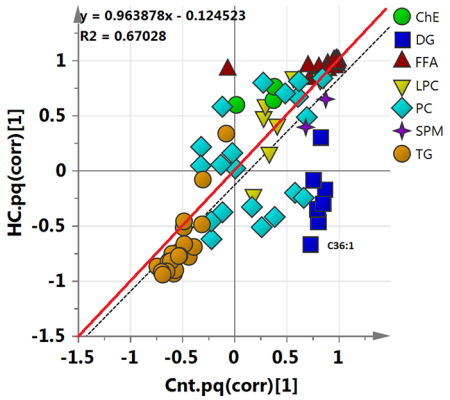


Figure 3. O2PLS SUS2 plot of control (Cnt) vs high cholesterol (HC); C36:1 equals diglyceride species containing two fatty acid chains consisting of a combined 36 carbon atoms with a single degree of unsaturation (e.g., 18:0 and 18:1). *p* represents the first component loadings from the *X* block, and *q* represents the first component loadings from the *Y* block. Lipid species nomenclature is as described in the text.

structural lipids and FFAs. TGs remained tightly grouped along the diagonal (highly correlated), but some PCs, DGs, and individual FFAs were more spread out (becoming less correlated) indicative of significant change between the dietary regimes. A SUS2 plot comparing LC and Cnt O2PLS models also revealed that TGs and FFAs were highly correlated (lie near to the diagonal) between both models and did not show large variation between diets. Conversely, PCs displayed low correlation between the two models (lie at extremes to the diagonal) and were hence influenced by changes in dietary regime. In Figure 3, C36:1 DG lies at the most extreme point from the diagonal and displays a high positive correlation in the control model and high negative correlation in the HC model. This observation was also prominent in the SUS2 plot of the LC versus HC models (Supporting Information Figure S10) where C36:1 DG is again located at an extreme distance to the diagonal. Interestingly, ChEs do deviate off the diagonal in Figure 3; however, they are present on the diagonal in the LC versus HC SUS2 plot. This observation suggests that ChE concentrations differ between the Cnt and LC/HC diets, but not between the LC and HC diets.

OPLS-DA lipidomics SUS results show both the DG and TG groups moving along the *Y* axis, shifting toward (increasing in) the HC model, suggesting that they play a larger role in hepatic metabolic response to high cholesterol intake.

Step 5: Biological Interpretation. Results of O2PLS SUS2 plot analyses showed that ChEs only change when comparing control and LC diets or control and HC diets. When comparing LC and HC models, ChEs rest squarely on the diagonal, indicating small changes between model variables. ChEs are an inert form of cholesterol, stored by the liver to prevent toxicity.[32] The SUS2 plots suggest that the livers of mice in both the LC and HC dietary regimes had reached a saturation point and could no longer store additional cholesterol esters. When this storage capacity is reached, the liver copes with the associated metabolic stress by intensifying the production and secretion of lipoproteins, which ultimately reach vascular walls and may cause atherosclerosis.[33]

The measurable effect of dietary LC and HC regimes in mice also appears to manifest itself in changes in specific carbon length PCs (observed in both OPLS-DA and O2PLS) and the large shift (correlation and concentration changes) of DGs along the *Y* axis of the O2PLS SUS2 plots, from their starting position near the FFAs to a different position near the TGs.

Interestingly, DGs were largely shifted off the diagonal in O2PLS SUS2 plots when comparing HC models with both control and LC models (Figure 3 and Supporting Information Figure S10). Examination of the O2PLS loadings plots confirmed that DGs dramatically shift (change correlations) in the HC-fed mice. It has been suggested that high cholesterol might induce ER stress (endoplasmic reticulum stress) and/or activation of inflammatory kinases.[34] Accordingly, the observed DG shifts could be related to metabolic attempts to mediate this response.[35] PCs also separate and change locale in both OPLS-DA SUS and O2PLS SUS2 plots in the LC and HC diets; however, cholesterol is more commonly associated with inflammation or stress, and it is less likely that PCs are a contributor as opposed to part of a reactionary response. However, it does not rule out the possibility that individual PCs are linked to induced liver stress.

The O2PLS model of the control FFA versus lipidomics data contained specific components that described uniqueness in two samples (4Cnt and 1Cnt), and the loading plot of these components suggested discrepancies in TG and FFA concentrations. In a similar manner, the O2PLS LC FFA versus lipidomics model showed uniqueness (from the specific component) in TG and PC profiles in two LC samples (1LC and 7LC). OPLS-DA control versus LC lipidomics, control versus HC lipidomics (orthogonal component) and O2PLS control FFA versus lipidomics models all showed unique variation in the same sample (4Cnt). The OPLS-DA control versus HC lipidomics models also showed unique variation in

the same sample (12Cnt), previously observable in the O2PLS yardstick analysis. After collating these specific/orthogonal components, three particular samples were identified that, while still possessing similar joint variation as the other samples, also showed significant unique variation (samples 4Cnt, 1LC, and 12Cnt). Interestingly, these three mice were also the largest in terms of body mass, weighing ∼4 g more than the group average. The O2PLS specific and OPLS-DA orthogonal components identified these characteristics and the underlying variables that caused them to be unique within their groups.

The multivariate analysis was repeated following removal of the six samples defining the original boundaries of biological variation. These O2PLS models proved to be similar to the original models; however, SUS2 plots comparing models contained greater spread between variables. Particularly, the TG cluster now spread across either side of the diagonal by a larger variance. The influence of TGs between dietary models was expected; TGs are commonly linked to the onset of hyperglycemia, a known risk factor for the onset of atherosclerosis.[36,37] Shifts in DGs were still observable, but their influence was lessened with the exception of DG (C36:1), which still had a large influence between diets in the Cnt, LC, and HC O2PLS models. The C36:1 DG species was previously found to increase ∼8-fold in the livers of mice on a western diet.[38]

O2PLS models samples based on their commonality; the samples (mice) displaying larger variation as a result of dietary regime were affected on a system wide basis, and hence should ideally remain part of the experimental set, representing the extreme conditions that a percentage of the sample population represent. OPLS-DA models maintained good predictability with these samples included; they did not interfere with class separation based on lipidomics profiles. Also of note, when removed from the experimental data, the samples are removed from the orthogonal components of the OPLS-DA models. The removal of samples that contribute to a large variable range or which may be outliers should be performed with caution and conducted with clear knowledge of what is biologically acceptable and relevant to a working hypothesis. These observations are a cautionary tale that even homogeneous transgenic animals can demonstrate unique biological responses.

Extension of Workflow to Therapeutic Intervention. The workflow was extended to analyze the effect of three atherogenic therapeutic interventions, fenofibrate (FF), rosuvastatin (RSV), and T-0901317 (T09), on the liver metabolism of mice fed a HC diet (derived from the same ApoE3Leiden data set[27]). Kleemann et al. reported that treatment with FF in HC-fed mice suppressed both acute and chronic atherosclerosis inflammatory processes.[28] This study also reported that FF affected the greatest number of biochemical pathways via the transcription factor PPAR-α (the primary drug target), which regulates hepatic lipoprotein metabolism. An O2PLS model generated from HC+FF FFA versus lipidomics data yielded a null (no) model. PPAR-α modulation resulted in shifts in intrahepatic lipid metabolism, but did not appear to cause shifts in FFA metabolism. Conversely OPLS-DA models on HC versus HC+FF for both FFA and lipidomics generated strong, robust models (high R^2 and Q^2), discriminating between samples of HC and HC+FF. The presence of these OPLS-DA models and absence of O2PLS models suggests that changes in FFA and in lipid metabolism are occurring through pathways existing beyond the measured variables in this data set.

The previous study reported that T09 and RSV mainly suppressed acute inflammatory response, each through unique biochemical pathways.[27] T09 primarily targets LXR-α and LXR-β nuclear receptors, while RSV is a structural inhibitor of 3-hydroxy-3-methyl-glutaryl coenzyme A reductase (HMG-CoA reductase), an enzyme that determines the rate-limiting production of hepatic cholesterol biosynthesis.

The FFA and lipidomics data for T09 (HC and HC+T09) produced both O2PLS and OPLS-DA models. An O2PLS SUS2 plot of HC+T09 versus HC and HC+T09 versus control (Supporting Information Figure S11) showed that the T09 drug had a net effect on DG metabolic profiles. Through the comparison of both SUS2 plots it was discerned that T09 causes DG profiles in HC mice to shift toward those typically observed in control mice. There is also an effect on TG and PC profiles, though the exact nature of interaction is hard to determine and remains sufficiently different from any of the existing diet models. ChEs also showed a return toward control metabolic profiles. LXR-α is known to directly impact linoleic acid and oleic acid metabolism (FFAs), which are components of the normal hepatic lipid metabolism. LXR treatment has been shown to quench liver inflammation and reduce serum amyloid A (SAA) and other liver-derived inflammation markers.[28,39]

RSV data generated an O2PLS model, lipidomics OPLS-DA model, but no FFA OPLS-DA model. The RSV lipidomics OPLS-DA model demonstrated a relationship between liver cholesterol metabolic changes as a response to the drug. O2PLS models of RSV+HC versus HC and RSV+HC versus control (Supporting Information Figure S11) showed that RSV has a net effect on DGs, returning them toward profiles typical of control mice. In line with the assumed link between DG and hepatic inflammation, statin treatment lowers liver-derived inflammation markers such as SAA and C-reactive protein (CRP).[40] ChE profiles, however, remained similar to those observed in HC mice. RSV was designed to specifically inhibit a single enzyme (HMG-CoA reductase) and thus cholesterol synthesis; the liver of the HC mouse was most likely saturated at the time of harvest; hence, ChE levels remained similar, but changes in DG metabolism manifested. The lack of an OPLS-DA FFA model may relate to RSV's drug specificity; it does not affect FFAs because there is low discriminate changes between HC FFA metabolites and HC+RSV FFA metabolites.

CONCLUSION

The proposed complementary application of PCA, OPLS-DA, and O2PLS in a systems biology workflow highlighted key variables affected by changes in dietary cholesterol in a murine model of atherosclerosis. Use of the SUS2 and SUS plots in combination with O2PLS and OPLS-DA models suggested that the livers of mice in both the LC and HC dietary regimes had reached a saturation point by the time livers were harvested and could no longer process cholesterol. This was expressed through changes in intrahepatic lipid (notably DG) concentrations. The modeling of the effect of three therapeutics on hepatic metabolism in HC-fed mice also highlighted DGs as metabolites of interest. In particular, O2PLS SUS2 plots showed that treatment with T09 or RSV returned the DG profile in HC-fed mice to that of control animals. The combination of dietary and therapeutic results suggests that shifts in intrahepatic DGs typifies the HC group, which reportedly develops atherosclerosis and atherogenesis-promoting inflammation in liver,[27] suggesting a role for DGs in

dx.doi.org/10.1021/ac301269r | Anal. Chem. 2012, 84, 7064−7071

eliciting chronic liver inflammation. This putative role is in agreement with reports of DGs being associated with nonalcoholic fatty acid liver disease.[38] This hypothesis requires testing but clearly demonstrates how the use of the proposed multivariate statistical workflow successfully integrated multiple data sets and led to the formation of a discrete testable hypothesis. Although these techniques were used in a complementary manner, only O2PLS was able to identify DGs as having a relationship with FFAs and, ultimately, a role in hepatic cholesterol responses. Accordingly, the use of PCA/OPLS alone would not have highlighted the potential role of DGs and especially the DG 36:1 species in atherogenesis or the effect of T09 or RSV on DGs in relation to dietary cholesterol.

■ ASSOCIATED CONTENT

Supporting Information

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: gemma.kirwan@gmail.com (G.M.K.); craig.wheelock@ki.se (C.E.W.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was supported by VINNOVA and JSPS under the Sweden−Japan Research Cooperative Program, the Åke Wibergs Stifelse, the Jeanssons Stiftelse, the Swedish Foundation for Strategic Research, Swedish Research Council (VR), Swedish national strategic e-science research program eSSENCE, and a VINNOVA VINN-MER International Grant. G.M.K. was supported by a JSPS postdoctoral fellowship, and C.E.W. was supported by a research fellowship from the Centre for Allergy Research.

■ REFERENCES

(1) Morrisson, N.; Cochrane, G.; Faruque, N.; Tatusova, T.; Tateno, Y.; Hancock, D.; Field, D. *OMICS* **2006**, *10*, 127−137.
(2) Hood, L. *Mech. Ageing Dev.* **2003**, *124*, 9−16.
(3) Kettaneh, N.; Berglund, A.; Wold, S. *Comput. Stat. Data Anal.* **2005**, *48*, 69−85.
(4) Trygg, J.; Wold, S. *Chemometrics* **2002**, *16*, 119−128.
(5) Shi, L. *Nat. Biotechnol.* **2010**, *28*, 827−838.
(6) Bylesjö, M.; Rantalainen, M.; Cloarec, O.; Nicholson, J. K.; Holmes, E.; Trygg, J. *Chemometrics* **2006**, *20*, 341−351.
(7) Gabrielsson, J.; Jonsson, H.; Airiau, C.; Schmidt, B.; Escott, R.; Trygg, J. *Chemometrics* **2006**, *20*, 362−369.
(8) Trygg, J. *Chemometrics* **2002**, *16*, 283−293.
(9) Skova, T.; Ballabiob, D.; Bro, R. *Anal. Chim. Acta* **2008**, *615*, 18−29.
(10) Wiklund, S.; Johansson, E.; Sjöström, L.; Mellerowicz, E. J.; Edlund, U.; Shockcor, J. P.; Gottfries, J.; Moritz, T.; Trygg, J. *Anal. Chem.* **2008**, *80*, 115−122.
(11) Bylesjö, M.; Eriksson, D.; Kusano, M.; Moritz, T.; Trygg, J. *Plant J.* **2007**, *52*, 1181−1191.
(12) Bylesjö, M.; Nilsson, R.; Srivastava, V.; Grönlund, A.; Johansson, A. I.; Jansson, S.; Karlsson, J.; Moritz, T.; Wingsle, G.; Trygg, J. *J. Proteome Res.* **2009**, *8*, 199−210.
(13) Richards, S. E.; Dumas, M.; Fonville, J. M.; Ebbels, T. M. D.; Holmes, E.; Nicholson, J. K. *Chemom. Intell. Lab. Syst.* **2010**, *104*, 121−131.
(14) Molteni, C. G.; Cazzaniga, G.; Condorelli, D. F.; Fortuna, C. G.; Biondi, A.; Musumarra, G. *QSAR Comb. Sci.* **2009**, *28*, 822−828.
(15) Musumarra, G.; Condorelli, D. F.; Fortuna, C. G. *Comb. Chem. High Throughput Screening* **2011**, *14*, 36−46.
(16) Ni, Y.; Su, M.; Lin, J.; Wang, X.; Qiu, Y.; Zhao, A.; Chen, T.; Jia, W. *Fed. Eur. Biochem. Soc., Lett.* **2008**, *582*, 2627−2636.
(17) Fonville, J. M.; Bylesjö, M.; Coena, M.; Nicholson, J. K.; Holmes, E.; Lindon, J. C.; Rantalainen, M. *Anal. Chim. Acta* **2011**, *705*, 72−80.
(18) Hedenström, M.; Wiklund, S.; Sundberg, B.; Edlund, U. *Chemom. Intell. Lab. Syst.* **2008**, *92*, 110−117.
(19) Lundstedt, T.; Hedenström, M.; Soeria-Atmadja, D.; Hammerling, U.; Gabrielsson, J.; Olsson, J.; Trygg, J. *Chemom. Intell. Lab. Syst.* **2010**, *104*, 112−120.
(20) Lundström, S. L.; Levänen, B.; Nording, M.; Klepczynska-Nyström, A.; Sköld, M.; Haeggström, J. Z.; Grunewald, J.; Svartengren, M.; Hammock, B. D.; Larsson, B. M.; Eklund, A.; Wheelock, Å. M.; Wheelock, C. E. *PLoS One* **2011**, *6*, e23864.
(21) Croixmarie, V.; Umbdenstock, T.; Cloarec, O.; Moreau, A.; Pascussi, J.; Boursier-Neyret, C.; Walthe, B. *Anal. Chem.* **2009**, *81*, 6061−6069.
(22) Fonville, J. M.; Richards, S. E.; Barton, R. H.; Boulange, C. L.; Ebbels, T. M. D.; Nicholson, J. K.; Holmes, E.; Dumas, M. *Chemometrics* **2010**, *24*, 636−649.
(23) Kim, Y.; Park, Y.; Yanga, S.; Kima, S.; Hyuna, S.; Cho, S.; Kim, Y.; Kwon, D.; Cha, Y.; Chae, S.; Choi, H. *Nutr. Res. (N.Y.)* **2010**, *30*, 455−461.
(24) Boccard, J.; Badoud, F.; Grata, E.; Ouertani, S.; Hanafi, M.; Mazerolles, G.; Lante, P.; Veuthey, J.; Saugy, M.; Ruda, S. *Forensic Sci. Int.* **2011**, *213*, 85−94.
(25) Bruce, S. J.; Breton, I.; Decombaz, J.; Boesch, C.; Scheurer, E.; Montoliu, I.; Rezzi, S.; Kochhara, S.; Guy, P. A. *J. Chromatogr., B* **2010**, *878*, 3015−3023.
(26) De Roover, K.; Ceulemans, E.; Timmerman, M. E. *Behav. Res. Methods* **2011**, *44*, 41−56.
(27) Kleemann, R.; Verschuren, L.; van Erk, M. J.; Nikolsky, Y.; Cnubben, N. H. P.; Verheij, E. R.; Smilde, A. K.; Hendriks, H. F. J.; Zadelaar, S.; Smith, G. J.; Kaznacheev, V.; Nikolskaya, T.; Melnikov, A.; Hurt-Camejo, E.; van der Greef, J.; van Ommen, B.; Kooistra, T. *Genome Biol.* **2007**, *8*, R200.
(28) Kleemann, R.; Bureeva, S.; Perlina, A.; Kaput, J.; Verschuren, L.; Wielinga, P. Y.; Hurt-Camejo, E.; Nikolsky, Y.; van Ommen, B.; Kooistra, T. *BMC Syst. Biol.* **2011**, *5*, 125.
(29) Rantalainen, M.; Cloarec, O.; Beckonert, O.; Wilson, I. D.; Jackson, D.; Tonge, R.; Rowlinson, R.; Rayner, S.; Nickson, J.; Wilkinson, R. W.; Mills, J. D.; Trygg, J.; Nicholson, J. K.; Holmes, E. *J. Proteome Res.* **2006**, *5*, 2642−2655.
(30) Eliasson, M.; Rännar, S.; Madsen, R.; Donten, M.; Marsden-Edwards, E.; Moritz, T.; Shockcor, J.; Johansson, E.; Trygg, J. *Anal. Chem.* **2012**, DOI: 10.1021/ac301482k.
(31) Zadelaar, S.; Kleemann, R.; Verschuren, L.; de Vries-Van der Weij, J.; van der Hoorn, J.; Princen, H. M.; Kooistra, T. *Arterioscler., Thromb., Vasc. Biol.* **2007**, *21*, 1706−1721.
(32) Greenberg, A. S.; Coleman, R. A.; Kraemer, F. B.; McManaman, J. L.; Obin., M. S.; Puri, V.; Yan, Q. W.; Miyoshi, H.; Mashek, D. G. *J. Clin. Invest.* **2011**, *121*, 2102−2110.
(33) Skålén, K.; Gustafsson, M.; Rydberg, E. K.; Hultén, L. M.; Wiklund, O.; Innerarity, T. L.; Borén, J. *Nature* **2002**, *13*, 750−754.
(34) Hotamisligil, G. S. *Nature* **2006**, *444*, 860−867.
(35) Van Rooyen, D. M.; Farrell, G. C. *Gastroenterol. Hepatol.* **2011**, *26*, 789−795.
(36) Geurian, K.; Pinson, J. B.; Weart, C. W. *Ann. Pharmacother.* **1992**, *26*, 1109−1117.
(37) Aronson, D.; Rayfield, E. J. *Cardiovasc. Diabetol.* **2002**, *1*, 1.
(38) Gorden, D. L.; Ivanova, P. T.; Myers, D. S.; McIntyre, J. O.; VanSaun, M. N.; Wright, J. K.; Matrisian, L. M.; Brown, H. A. *PLoS One* **2011**, *6*, e22775.
(39) Verschuren, L.; de Vries-van der Weij, J.; Zadelaar, S.; Kleemann, R.; Kooistra, T. *J. Lipid Res.* **2009**, *50*, 301−311.

(40) Kleemann, R.; Princen, H. M. G.; Emeis, J. J.; Jukema, J. W.; Fontijn, R. D.; Horrevoets, A. J. G.; Kooistra, T.; Havekes, L. M. *Circulation* **2003**, *108*, 1368−1374.