analytical chemistry

Article

pubs.acs.org/ac

OnPLS-Based Multi-Block Data Integration: A Multivariate Approach to Interrogating Biological Interactions in Asthma

Stacey N. Reinke,[*,†,‡] Beatriz Galindo-Prieto,[§,∥,⊥] Tomas Skotare,[§] David I. Broadhurst,[‡] Akul Singhania,[#,▽] Daniel Horowitz,[○] Ratko Djukanović,[#,◆] Timothy S.C. Hinks,[#,◆,††] Paul Geladi,[‡‡] Johan Trygg,[§] and Craig E. Wheelock[*,†,§§]

[†]Division of Physiological Chemistry 2, Department of Medical Biochemistry and Biophysics, Karolinska Institute, SE-171 77 Stockholm, Sweden

[‡]Centre for Integrative Metabolomics and Computational Biology, School of Science, Edith Cowan University, Perth 6027, Australia

[§]Computational Life Science Cluster, Department of Chemistry (KBC) and [∥]Industrial Doctoral School (IDS), Umeå University, SE-901 87 Umeå, Sweden

[⊥]Department of Engineering Cybernetics (ITK), Norwegian University of Science and Technology (NTNU), 7491 Trondheim, Norway

[#]Clinical and Experimental Sciences, University of Southampton Faculty of Medicine and [◆]NIHR Southampton Respiratory Biomedical Research Unit, Southampton University Hospital, Southampton SO16 6YD, U.K.

[▽]Laboratory of Immunoregulation and Infection, The Francis Crick Institute, London NW1 1AT, U.K.

[○]Janssen Research and Development, High Wycombe HP12 4DP, Buckinghamshire, U.K.

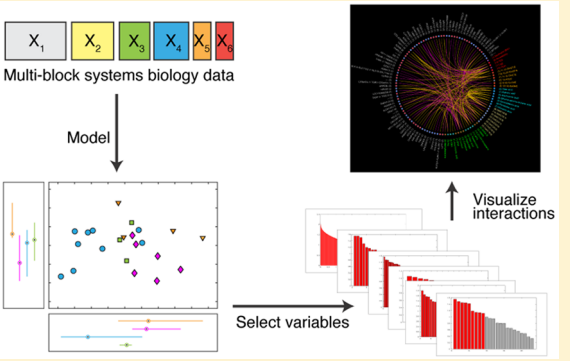
[††]NIHR Oxford Biomedical Research Centre/Respiratory Medicine Unit, NDM Experimental Medicine, University of Oxford, Level 7, John Radcliffe Hospital, Oxford OX3 9DU, U.K.

[‡‡]Forest Biomass and Technology, Swedish University of Agricultural Sciences, SE 90183 Umeå, Sweden

[§§]Gunma University Initiative for Advanced Research (GIAR), Gunma University, Maebashi 371-8510, Japan

S *Supporting Information*

ABSTRACT: Integration of multiomics data remains a key challenge in fulfilling the potential of comprehensive systems biology. Multiple-block orthogonal projections to latent structures (OnPLS) is a projection method that simultaneously models multiple data matrices, reducing feature space without relying on a priori biological knowledge. In order to improve the interpretability of OnPLS models, the associated multi-block variable influence on orthogonal projections (MB-VIOP) method is used to identify variables with the highest contribution to the model. This study combined OnPLS and MB-VIOP with interactive visualization methods to interrogate an exemplar multiomics study, using a subset of 22 individuals from an asthma cohort. Joint data structure in six data blocks was assessed: transcriptomics; metabolomics; targeted assays for sphingolipids, oxylipins, and fatty acids; and a clinical block including lung function, immune cell differentials, and cytokines. The model identified seven components, two of which had contributions from all blocks (globally joint structure) and five that had contributions from two to five blocks (locally joint structure). Components 1 and 2 were the most informative, identifying differences between healthy controls and asthmatics and a disease−sex interaction, respectively. The interactions between features selected by MB-VIOP were visualized using chord plots, yielding putative novel insights into asthma disease pathogenesis, the effects of asthma treatment, and biological roles of uncharacterized genes. For example, the gene *ATP6 V1G1*, which has been implicated in osteoporosis, correlated with metabolites that are dysregulated by inhaled corticoid steroids (ICS), providing insight into the mechanisms underlying bone density loss in asthma patients taking ICS. These results show the potential for OnPLS, combined with MB-VIOP variable selection and interaction visualization techniques, to generate hypotheses from multiomics studies and inform biology.



In the postgenomic era, data-driven science has become increasingly necessary because of the vast array of instrumentation that is capable of generating thousands of data points for a single analytical observation.[1,2] In addition to

Received: July 17, 2018
Accepted: October 18, 2018
Published: October 18, 2018

DOI: 10.1021/acs.analchem.8b03205
Anal. Chem. XXXX, XXX, XXX−XXX

using classical univariate statistical methods, machine-learning techniques have become routinely used to interrogate and understand vast amounts of data.[3,4] Two common characteristics of -omics data are that the number of measured variables is vastly greater than the number of observations[5] and that there is a degree of multicollinearity between variables.[6] As such, computational methods that project high dimensional data into a smaller number of component variables have become commonplace.[7] Multivariate projection methods such as principal components analysis (PCA),[8] partial least squares discriminant analysis (PLS-DA),[9] and canonical variate analysis (CVA),[8] together with hierarchical cluster analysis (HCA),[10] random forests,[11] and support vector machines (SVM),[12] are all used to analyze -omics data.[3,4] PLS-DA and its extension, orthogonal projection to latent structures discriminant analysis (OPLS-DA),[13,14] have become popular projection methods in the metabolomics community.[15] As modeling methods become increasingly complicated, they have also become concomitantly difficult to interpret. Assignment of the variable importance often becomes an a posteriori statistical process based on either permutation testing or random resampling (e.g., confidence intervals derived from bootstrap/jackknife statistics).[16] For methods based on a PLS algorithm, the direct statistical method of variable influence on projection (VIP)[17,18] is often used to estimate variable contribution to the resulting models.

In recent years, as the -omics sciences have matured, it has become common to acquire data from multiple -omics platforms in a single biological experiment. As such, each biological sample is interrogated by multiple analytical platforms, which in turn can be linked to multiple sources of experimental metadata. Data from each platform (or measurement context) can be considered a discrete *block*, with multiple blocks making up the complete data set of the experiment. Multivariate projection methods such as OPLS-DA have proven successful in modeling the underlying latent biological structure within a single high dimensional data block; however, they are theoretically unsuitable for modeling multiple data blocks simultaneously. There are two reasons for this issue. First, if multiple data blocks are concatenated into a single matrix, with no accounting for measurement context, then the subsequent model can be considered as a single projection model, where the weighting of each variable is governed by the total sum of squares.[19] This, in principle, demands that each block is normalized to the same size, to avoid a projection model that is biased toward the impact of the data set with the most variables. In practice, this can be problematic, particularly when there are a mix of blocks of vastly different sizes. For example, in a model concatenating 20 000 transcripts, 200 metabolites, and 20 clinical variables, the transcripts would over-represent the global data structure and thus have a larger contribution to the resulting model. In multi-block modeling, this is not an issue, as each block is treated independently. This approach leaves flexibility to scale individual variables according to importance and also to keep variables in their original unit. Second, each individual data set is associated with its own underlying structure,[19,20] describing the true biological variance and also platform-specific measurement error. Covariance of biological latent structure across multiple data blocks is implicit; however, it is a fair assumption that the measurement error across multiple blocks will be independent and thus easily ignored at this block-interaction level. Conversely, if multiple data blocks are concatenated into a single data set before projection, the model will struggle to effectively separate true biological structure from block-specific noise and result in erroneous interpretation of the conglomerate projection model.

To address the need for multivariate methods to simultaneously model multiple data matrices, a number of multi-block data integration methods have been proposed.[21−23] In 2011, Löfstedt and Trygg[24] proposed a novel multi-block multivariate method called OnPLS, which utilizes the framework of OPLS to decompose data from more than two input matrices. Multi-block models, such as OnPLS, are fully symmetric, meaning each data block is weighted to allow an equal contribution to the model, regardless of the number of variables or underlying data structure within each block.[25] Multi-block approaches offer further advantages over single block or block concatenation in biomarker discovery. First, the validity of any true biological biomarker is significantly increased if there is a clear covariance between data blocks, thus reducing the possibility of false discovery.[26] Second, contrary to block-concatenation modeling, which is strongly biased toward the globally joint variation, multi-block analysis decomposes the different levels of variation (global, local, unique)[27] such that relatively small but informative trends are also identified. Recently, Galindo-Prieto et al. adapted the VIP concept for multi-block data analysis (multi-block variable influence on orthogonal projections method,[28] MB-VIOP) to identify the variables that contribute to these different levels of joint structure.

The aim of this study was combine OnPLS and MB-VIOP with data visualization methods to create a workflow capable of simultaneously modeling and investigating interactions between multiple -omics data blocks. The study chosen for this purpose was a subset from a previously reported asthma cohort, for which multiple -omics data sets were acquired in isolation.[29,30] These analyses included untargeted metabolomics, targeted metabolite assays, differential immune cell population analyses, and cytokine arrays. Additionally, for the present study, transcriptomics of peripheral blood T cells was performed. OnPLS modeling and MB-VIOP were then used to integrate the disparate data blocks into a single model, which was then interrogated to identify novel interactions between the data blocks and disease status as well as other clinical end points.

EXPERIMENTAL SECTION

Clinical Cohort. Briefly, 12 healthy controls and 10 severe asthmatics were included from the original study.[29] Transcriptomics was subsequently performed on peripheral blood T cells, and metabolomics/metabolic profiling assays were performed on serum. All participants were enrolled from the NIHR Southampton Respiratory Biomedical Research Unit and University Hospital Southampton outpatient clinics; all provided written informed consent. The National Research Ethics Service Committee South Central—Southampton B ethics committee (UK; ref 10/H0504/2) approved this study. Clinical classification and enrollment criteria were previously described.[29,31] Participant data were included in the present study if they were classified as either healthy control or severe asthmatic individuals in the existing cohort, and data from all data blocks (described in next section) were collected.

Sample Collection and Analyses. Details of sample collection and transcriptomics analyses are available in the Supporting Information. Details of analytics, quality control,

DOI: 10.1021/acs.analchem.8b03205
Anal. Chem. XXXX, XXX, XXX−XXX

and data cleaning for metabolomics, targeted metabolic assays, and clinical assays were performed as previously described.[29,30]

Data Blocks and Processing. Six data blocks were used for modeling: Transcriptomics, Sphingolipids, Metabolomics, Fatty Acids, Oxylipins, and Clinical Data (Figure 1). A

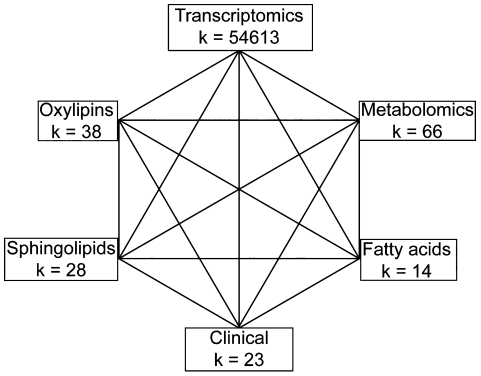


Figure 1. Schematic of potential shared structure between data blocks. The six data blocks used in this study are shown with their respective numbers of variables. The diagram shows all possible shared structure connections between the data blocks.

complete list of all variables included for each data block is provided in Tables S1−S6. The data blocks were defined by a priori knowledge about both the system under observation and the measurement technology.[19] The primary consideration was that the underlying structure of the data could possibly confound the biological interaction between blocks. To avoid bias in combining information from different probes for one gene, all non-QC probes were included for OnPLS modeling; the Transcriptomics block included 54 613 variables. This approach is commonly used for analyzing transcriptomics data.[32] Four data blocks represented serum metabolites: Sphingolipids (28 variables, targeted assay), Metabolomics (66 variables, untargeted assay screened against an in-house chemical library), Fatty Acids (14 variables, targeted assay), and Oxylipins (38 variables, targeted assay). A total of 23 clinical variables were combined into the Clinical data block; these variables were derived from typical clinical assays and measurements and included lung function tests, bronchoalveolar lavage fluid and peripheral blood T cell populations, serum cytokines, and serum vitamin D3. For clinical data, values that were missing due to being below the limit of detection (LOD) of the respective assay were imputed with 1/10 of the lowest measured value, because the LOD was not known for each assay, and OnPLS cannot process missing values. Data that were missing for an entire subset array of the clinical data (e.g., for individuals missing the cytokine assay) were imputed using the median value of the corresponding clinical group (control or asthma). Remaining missing values were replaced using PCA imputation. Prior to OnPLS model calculation, all data (except for transcriptomics) were log-transformed. All data were then scaled to unit variance.

OnPLS Model Calculation and Visualization. The OnPLS model simultaneously analyzed the data matrices, returning output matrices of shared information (components), as described.[27] These output matrices reveal shared data structure on three levels for each data matrix, which can be summarized as

$$X_i = \underbrace{X_G}_{\text{globally joint}} + \underbrace{X_L}_{\text{locally joint}} + \underbrace{X_U}_{\text{unique}} + \underbrace{E}_{\text{residual noise}} \quad (1)$$

Globally joint components reveal structure that is shared by all input data matrices. Locally joint components reveal structure shared by two or more, but not all, of the input matrices. Finally, unique components identify latent structure that is present in only one input matrix. The OnPLS model returned separate score vectors for each data block in each component. To identify the sources of biological variance explained by the OnPLS components, the component scores for each block were correlated with metadata variables not included in the clinical data block: clinical class (control vs asthma), sex, age, BMI, dose of inhaled and oral corticosteroids, and smoking (current/former smoker vs never smoked). The resulting Pearson correlation coefficients were presented as a metadata correlation plot.[33] To visualize the overall OnPLS model, hierarchical principal component analysis (PCA)[34] was used to summarize the 30 OnPLS score vectors, resulting in 2 PCA components describing the relationships in the OnPLS model. Prior to calculating the PCA model, the score vectors were scaled to unit variance. The PCA score plot showed individual participants, and the loadings plot displayed the score vectors from the OnPLS model, labeled by block type and OnPLS model component number.

MB-VIOP Concept, Motivation, and Calculation. Multi-block variable influence on orthogonal projections (MB-VIOP) is a feature selection method that (i) sorts the input variables by importance for data interpretation in OnPLS models, either for the total model (all variation types together) or per component (global, local, or unique variations separately), and (ii) explores the connections between the variables (either in the same or a different data matrix) that contribute to explain the same component (latent variable) in the multi-block system. Multi-block-VIOP is a model-based variable selection method, because it uses the n preprocessed data matrices, the score vectors, and the normalized loading vectors from an OnPLS model. OnPLS regression can relate the data matrices according to the model component; however, it must be emphasized that not all input variables of these related matrices will connect among themselves to explain the variation contained in a specific model component. The MB-VIOP algorithm is necessary to sort the input variables according to their connections for interpreting the variation contained in one or more specific components. Furthermore, MB-VIOP finds the degree of importance of each variable in the correct proportion for a multi-block system, which cannot be achieved by the OnPLS normalized loadings plot.[35]

The calculation of the MB-VIOP values can be summarized as the Hadamard products of the normalized loadings multiplied by the ratio of the variation explained by a model component and the cumulated variation. After a block- and component-wise iterative algorithm with all input variables from the six data matrices involved, the resulting MB-VIOP vectors were normalized by Euclidean norm and by the number of original (input) variables raised to the 1/2 power. The variables of interest that were identified by MB-VIOP were selected as a subset for further multivariate analysis as shown below. For additional details about the MB-VIOP fundamentals and algorithm, readers are referred to the original references.[28]

Data Visualization. The between-block covariance of the subset of variables contributing to Components 1 and 2 of the OnPLS model were visualized using chord plots.[36] Using the

DOI: 10.1021/acs.analchem.8b03205
Anal. Chem. XXXX, XXX, XXX−XXX

variables reaching a defined MB-VIOP threshold, a chord plot was constructed by first calculating the Spearman rank correlation coefficient (r) for each pairwise combination of variables with MB-VIOP values above a threshold. Those variables where a significant ($p < 0.001$) between-block correlation existed were presented as nodes in a circle (grouped by block), and the correlation represented as a colored arc (yellow being a positive correlation and purple a negative correlation). The number of arcs associated with a given node is recorded in parentheses next to the name of the variable. Each chord plot was constrained such that within-block correlations were ignored.

Data modeling (OnPLS), variable selection (MB-VIOP), a posteriori analyses, and creation of plots were performed using MATLAB 2018a (Mathworks, Natick, MA, USA). Correlation coefficients for the metadata correlation plots were calculated using functions from SciPy (http://www.scipy.org/), and the plot was created using the Matplotlib library.[37] SIMCA v15 (Umetrics, Umeå, Sweden) was used to perform OPLS-DA analysis.

RESULTS AND DISCUSSION

Study Population. A total of 22 participants from a previously described cohort[29,30] were included in this study (12 healthy control individuals and 10 individuals with severe asthma). Clinical information is presented in Table 1. Age and BMI were significantly higher in the severe asthmatic group and thus represented confounders in the study. Furthermore, all individuals in the severe asthmatic group were treated with inhaled and/or oral corticosteroids (ICS/OCS). Although the sex ratio and proportion of smokers were also different, they were not significantly altered between the two groups.

Table 1. Clinical Data

	healthy control (N = 12)	severe asthma (N = 10)
age (years)	26.5 (24.8, 30.8)	63 (43.5, 63)
sex (M/F)	9/3	4/6
BMI (kg/m^2)	24.1 (22.6, 43.2)	34.0 (27.4, 43.2)
smoking status		
never smoker (#)	11	6
current/former smoker (#)	1	4
treatment		
inhaled corticosteroids (#, median dose[b])	0	10 (1280)
oral corticosteroids (#)	0	3

[a]Values are medians (interquartile range) or numbers. [b]Beclomethasone dipropionate equivalent μg.

OnPLS Model. The OnPLS model calculated seven components that shared joint structure between at least two of the data blocks (Table 2). Two components (1 and 4) had globally joint structure, with contributions from all six blocks. The remaining components had locally joint structure, with between 2 and 5 data blocks contributing to the joint structure. The model did not identify any unique components.

The amount of variance explained in each component, for each data block, as well as cumulative variance explained by the model is reported in Table 2. Only 37% of the total variance in the Transcriptomics data block was explained, indicating that the majority of the information contained in this block is not descriptive for describing asthma. This could be due to the global and unbiased nature of the platform and/or the fact that the transcriptomics data were derived from the entire peripheral blood CD3+ T cell population. It would be of more clinical relevance to target specific cell subpopulations in a single-cell transcriptomics approach.[38] The clinical data described only 55% of the variance in the Clinical block; however, 16 of the 22 variables were either differential immune cell counts/subpopulation frequencies or cytokines produced by immune cells. Given the pathophysiological heterogeneity of asthma, traditional cell population and cytokine measures alone are insufficient to describe the disease.[39]

The OnPLS model explained >70% of the variance in each metabolic profiling data block (Sphingolipids, Fatty Acids, and Oxylipins) and 56% of the Metabolomics block. This higher degree of explained variance can be attributed to the selective association between these variables and asthma. These targeted assays were performed to confirm findings from the initial metabolomics screen.[30] While not all targeted metabolites were originally detected using metabolomics, they represent biological processes known to be involved in inflammation. This point is of particular relevance in that it is not the number of variables in a given data block that is the primary driver but rather the inherent biological content.[4,9] This facet makes it meaningful to combine disparate -omics blocks of varying structure into a single OnPLS model and demonstrates the utility of this approach for data modeling. However, there is the expected caveat that data blocks that contain higher levels of biological structure will have a concomitant increase in contribution to the overall OnPLS model.

To determine the biological factors associated with each OnPLS component and data block, model score vectors were correlated with a number of known biological factors (Figure 2). Component 1 scores from all blocks positively and significantly ($p < 0.05$) correlated with disease status (healthy vs asthma), age, and BMI. All blocks, except Fatty Acids, positively and significantly ($p < 0.05$) correlated with ICS and OCS dose. Transcriptomics, Fatty Acids, and Clinical scores

Table 2. OnPLS Model Summary

component	connection	Transcriptomics	Sphingolipids	Metabolomics	Fatty Acids	Oxylipins	Clinical
1	global	8%	33%	16%	25%	8%	17%
2	local	7%	11%	7%	-	40%	15%
3	local	-	14%	-	-	7%	-
4	global	8%	9%	10%	39%	7%	14%
5	local	6%	6%	9%	-	-	-
6	local	4%	13%	7%	-	-	-
7	local	5%	-	7%	10%	9%	9%
sum		37%	86%	56%	74%	71%	55%

DOI: 10.1021/acs.analchem.8b03205
Anal. Chem. XXXX, XXX, XXX−XXX

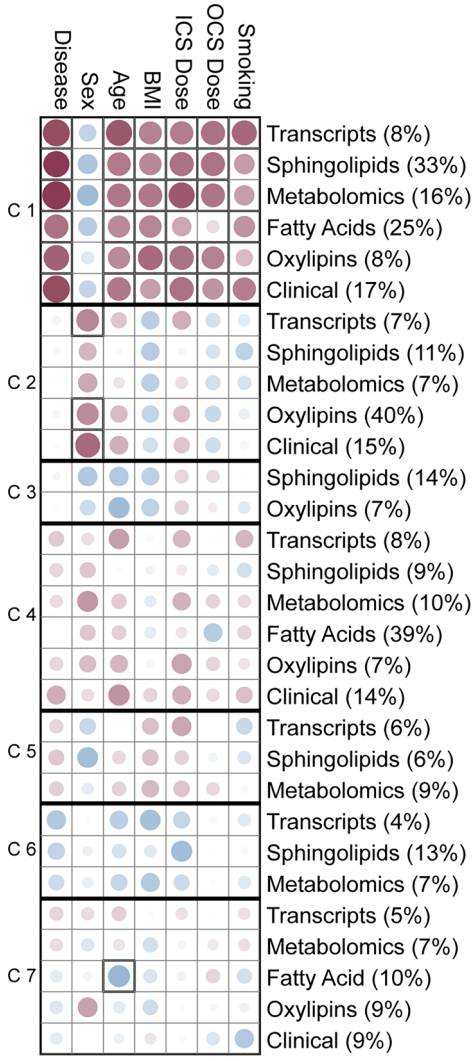


Figure 2. Correlation between model scores and metadata. Circle size and color intensity are proportional to strength of correlation (larger and darker indicates strong correlation). Red, positive correlation; blue, negative correlation. Thick outline around box, significant correlation ($p < 0.05$). The amount of variance that is explained by each data block, in each component, is shown in parentheses. Components are listed as C1−C7 on the left side of the figure.

correlated with smoking status (nonsmoker vs has ever smoked). As expected, age, BMI, and corticosteroid treatment were all confounded with disease status (Table 1); thus, explained variation in the model because of these factors was not distinguished from that of disease. The Component 2 scores for the Transcriptomics, Oxylipins, and Clinical blocks significantly ($p < 0.05$) and positively correlated with sex. While sex was not a significant confounder in this study, the distribution between the two classes was different. This highlights the utility for OnPLS to identify biological sources for variation in -omics data. The scores for Components 3−6 did not correlate significantly with any of the listed biological factors and likely describe either a combination of recorded biological factors or biological factors that were either not observed or not recorded in this study. As such, this highlights the importance of strict experimental design measures and extensive record keeping in data-driven sciences. Despite being a confounder in the study, age negatively correlated with Component 7 Fatty Acid scores and highlights the potential for OnPLS to identify underlying biology associated with data blocks.

PCA of OnPLS Score Vectors. To visualize the entire OnPLS model, principal components analysis (PCA) was performed on the scaled OnPLS score vectors (hierarchical PCA, Figure 3). The first principal component (PC1) showed

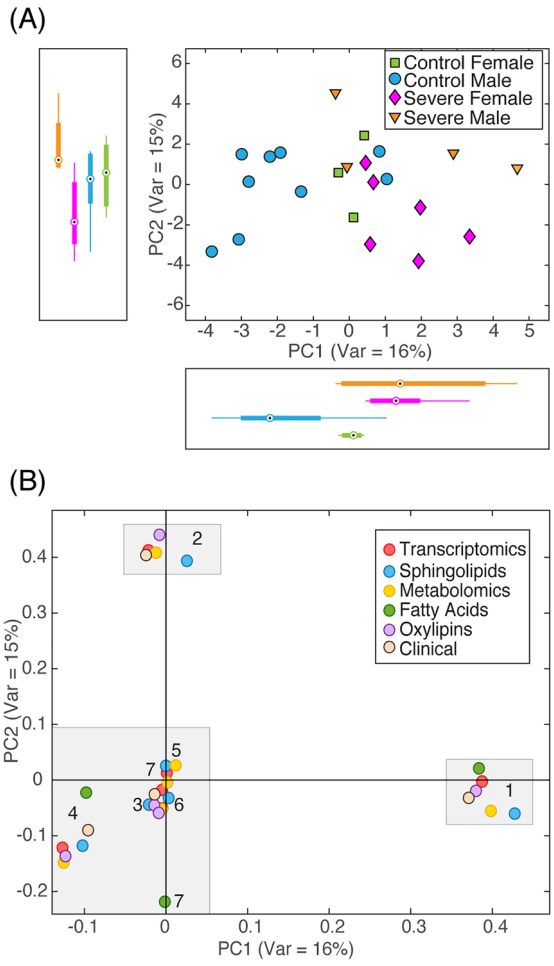


Figure 3. PCA visualization of OnPLS model score vectors. Score vectors from the OnPLS model were scaled to unit variance before performing H-PCA. (A) Score plot. Green squares, control females; blue circles, control males; purple diamonds, severe females; orange inverted triangles, severe males. Bar graphs on axes show distribution of each group along the respective axis. (B) Loadings plot. Red, Transcriptomics; blue, Sphingolipids; yellow, Metabolomics; green, Fatty Acids; purple, Oxylipins; tan, Clinical. Numbers represent OnPLS components, from which score vectors originate. Shaded boxes are for visualization purposes only.

a separation between healthy controls and asthmatic individuals in the score plot (Figure 3A). Aligning with the results of the correlation analysis, this separation was driven by the OnPLS Component 1 score vectors (Figure 3B). It was then expected that PC2 would solely describe a sex difference, as OnPLS Component 2 score vectors drove the separation. Interestingly, PC2 actually described an interaction between disease and sex (Figure 3A). While there was a sex difference among asthmatics, this was not observed in the controls. Investigating the interaction between sex and disease was not an aim of the original cohort study; however, this interaction was identified by simultaneously modeling all the data in

DOI: 10.1021/acs.analchem.8b03205
Anal. Chem. XXXX, XXX, XXX−XXX

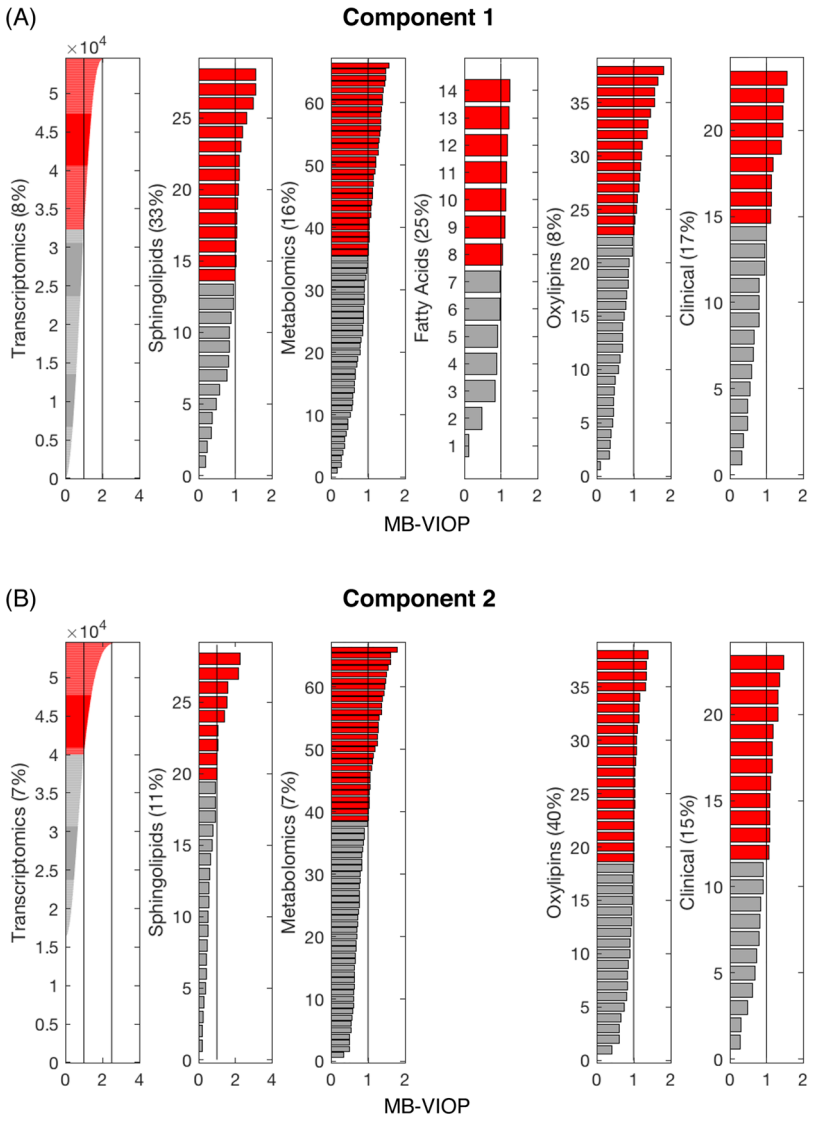


Figure 4. MB-VIOP variable selection for OnPLS Components 1 and 2. The MB-VIOP values are shown for each block in Components 1 and 2. Gray bars, variables with MB-VIOP ≤ 1.0; red bars, variables with MB-VIOP > 1.0. Vertical lines are drawn to show MB-VIOP > 1.0 threshold for all blocks in addition to the increased MB-VIOP thresholds of >2.0 and >2.5 for Transcriptomics in Components 1 and 2, respectively. Percentages reflect the amount of variance described by each component, for each data block. (A) Component 1. (B) Component 2.

combination with integrative visualization. In addition, the hierarchical PCA model corroborates the correlation analysis, showing that OnPLS Components 1 and 2 contain the most structural information. Therefore, these components were selected for further exploration with MB-VIOP analysis.

Multi-block Variable Influence on Orthogonal Projections (MB-VIOP). To further investigate the variables and their interactions underlying the shared structure of OnPLS components 1 and 2, MB-VIOP variable selection and subsequent correlation analysis were applied. A MB-VIOP threshold of >1.0 was used to select the variables of interest from each component. For Component 1, 22 297 transcripts, 31 metabolites, 15 sphingolipids, 7 fatty acids, 16 oxylipins, and 9 clinical variables contributed to explaining the shared structure describing disease separation (Figure 4A). For visualization purposes, the MB-VIOP threshold was increased to 2.0 for the Transcriptomics data block, leaving 151 variables. For Component 2, 14 618 transcripts, 28 metabolites, 9 sphingolipids, 20 oxylipins, and 12 clinical variables contributed to explaining the shared structure describing the interaction between sex and disease (Figure 4B). The Transcriptomics block appeared to have a strong influence on the disease−sex interaction, with 1487 transcripts passing the higher MB-VIOP threshold of 2.0; thus, the threshold was further increased to >2.5 to identify only the strongest contributions, leaving 203 transcripts. The complete list of all MB-VIOP values calculated for Components 1 and 2 is presented in Tables S1−S6.

In order to identify between-block biological interactions in Components 1 and 2, chord plots were used to visualize correlations of variables passing the specified MB-VIOP thresholds (Figure 5). This approach revealed a number of interesting interactions, of which a selected few are discussed as examples of the application of the proposed workflow. Five metabolites that correlated with ICS dose[30] (cortisol; cortisone; dehydroepiandrosterone sulfate, DHEA-S; N-palmitoyltaurine, pipecolate) passed the MB-VIOP threshold criteria for Component 1. These metabolites correlated with the transcripts of 21 unique genes (Figure 5A), of which *ATP6 V1G1* was particularly interesting. *ATP6 V1G1* has been

DOI: 10.1021/acs.analchem.8b03205
Anal. Chem. XXXX, XXX, XXX−XXX

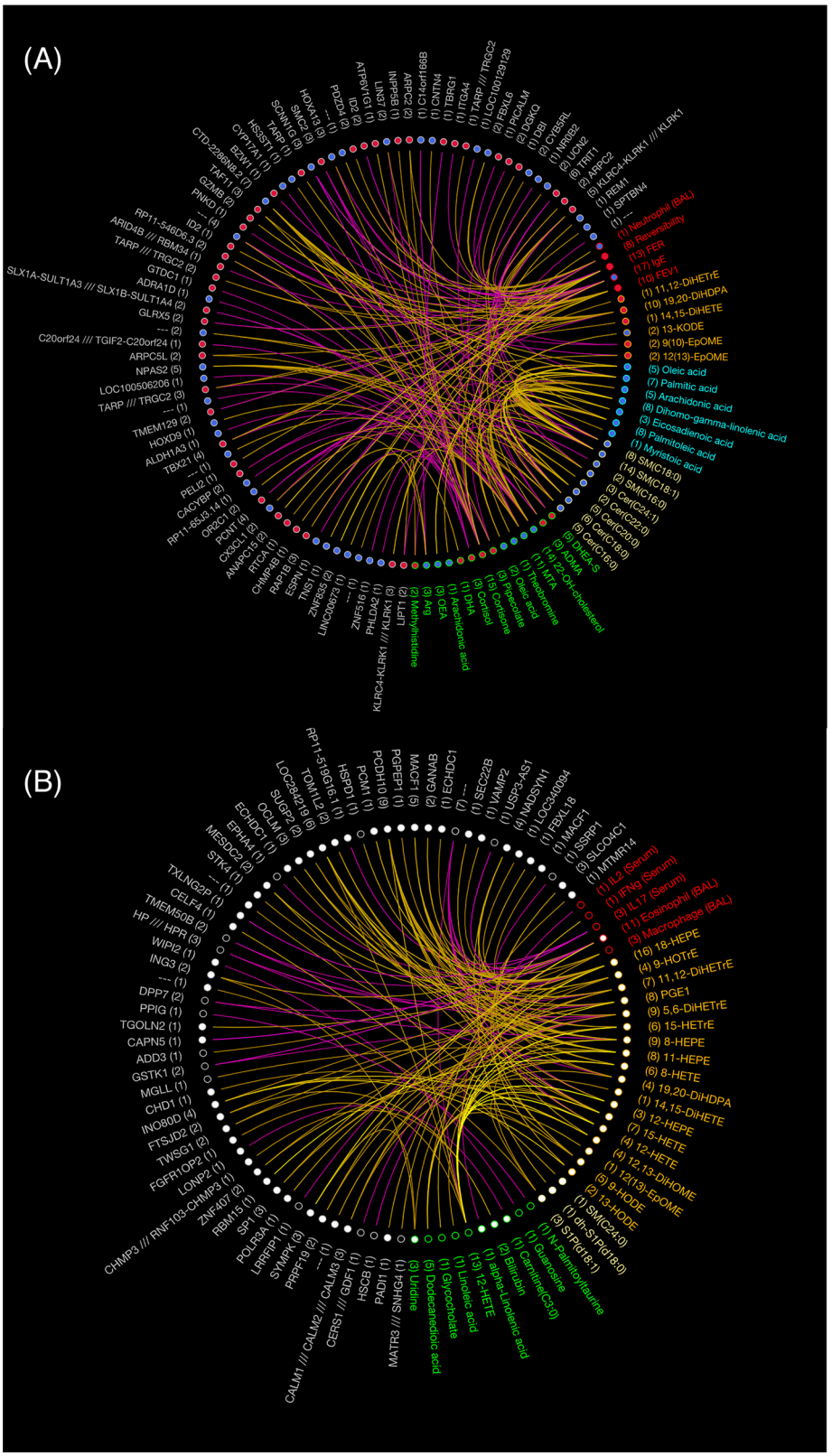


Figure 5. Chord plots showing between-block correlations. (A) Component 1. (B) Component 2. Chord plots were made by calculating the Spearman rank correlations for each pairwise comparison of variables meeting the MB-VIOP thresholds. Variables with a significant ($p < 0.01$) between-block correlation were presented in the chord plots. Nodes represent variables. Text color is associated with block: gray, Transcriptomics; green, Metabolomics; yellow, Sphingolipids; blue, Fatty Acids; orange, Oxylipins; red, Clinical. The number of correlations associated with a given node is noted in parentheses next to the name of the variable. Node color represents direction of change. Component 1: blue, increased in asthma; red, decreased in asthma. Component 2: white, increased in females; black, increased in males. Chords represent correlations: yellow, positive correlation; purple, negative correlation. Each chord plot was constrained such that within-block correlations were ignored. (---) denotes noncoding gene transcripts.

DOI: 10.1021/acs.analchem.8b03205
Anal. Chem. XXXX, XXX, XXX−XXX

implicated in osteoporosis and specifically osteoclast function,[40] which is a known side-effect of ICS treatment.[41] This novel link may provide insights to the mechanisms underlying bone density loss in asthma patients taking ICS. In addition, *NPAS2*, a transcription factor involved in mediating circadian rhythm,[42] correlated with five metabolites, four of which were ceramides (Figure 5A). Evidence suggests that ceramide levels fluctuate diurnally;[43,44] however, to our knowledge, this is the first time an association has been made between *NPAS2* and ceramides. More importantly, as all samples were collected at the same time of day (between 09:00 and 11:00), this supports emerging evidence of dysregulated circadian rhythm gene expression in asthma.[45] Indeed, experiencing symptoms more than once per week was a classification criterion of severe asthma.[29] The disease−sex interaction identified by Component 2 was largely driven by differential bronchoalveolar lavage cell profiles (eosinophils, macrophages) and oxylipins (Figure 5B). It also identified a high degree of correlation between the oxylipins and both *PCDH10* and the uncharacterized gene locus LOC284219, suggesting that these genes may play a previously unidentified role in oxylipin metabolism. Together, these examples highlight the value of this method for interrogating biology and generating hypotheses from multiomics data.

By combining OnPLS multi-block modeling with MB-VIOP variable selection and various visualization methods, the composite of data derived from this study could be interrogated. Where methods such as OPLS are useful for identifying covariance in isolated data blocks, OnPLS offers the advantage of identifying combined covariance, thus offering a more complete understanding of the whole system. For example, when OPLS was applied to the Metabolomics data block in isolation, 21 variables had a $VIP_{OPLS} > 1.0$ with dehydroepiandrosterone-sulfate (DHEA-S) being the strongest driver of the control−asthma difference (Supplemental Tables). Component 1 of OnPLS had 31 variables with a MB-VIOP > 1.0, 15 of which were unique to OnPLS modeling. Whereas DHEA-S was a major driver in the covariance in the single-block analysis, it was less important in the combined covariance of the OnPLS model. The Transcriptomics, Oxylipin, and Clinical data blocks showed similar trends, with OPLS and OnPLS revealing different biological insights (data not shown).

While the present study shows the potential for OnPLS-based modeling to be useful for simultaneously modeling multiple data blocks and generating hypotheses, it is limited by sample size and study power. Furthermore, OnPLS is currently unable to derive a block weighting such that MB-VIOP values can be scaled and directly compared across all blocks. Accordingly, MB-VIOP values can only be directly compared within a given data block and not between blocks. In interpreting the results, one must consider the overall contribution, not only of the block per se but also of the individual variables, to the respective component.

CONCLUSIONS

The multi-block OnPLS method combined with MB-VIOP variable selection and interaction visualization techniques yielded putative novel insights into asthma disease pathogenesis, the effects of asthma treatment, and biological roles of genes. The current study was performed in a worst-case scenario approach using a small sample set, with unbalanced groups and multiple study confounders. While these issues limit the ability of the different components of the OnPLS model to identify unique biological sources of variation, it demonstrates the potential for this method for identifying key structure in -omics data integration. It is likely that in large well-designed studies, the different components would be able to identify and explain other sources of biological and/or experimental variability (e.g., therapeutics, center bias, diet). It is also possible that this approach would be useful in identifying subphenotypes of disease, with different subgroups and/or mechanisms described by different components. We therefore propose that OnPLS modeling can be incorporated into large-scale molecular phenotyping studies for stratified medicine. Given that the -omics technologies detect molecules that function in a highly interdependent and dynamic manner within a living system, multi-block methods such as OnPLS, together with MB-VIOP and interaction visualization, provide a logical approach to investigating systems biology.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.8b03205.

Supporting methods for transcriptomics (PDF)

$MBVIOP_{OnPLS}$ and $VIP_{OPLS-DA}$ values for Block 1 (Transcriptomics) variables, Table S1; $MBVIOP_{OnPLS}$ and $VIP_{OPLS-DA}$ values for Block 2 (Sphingolipids) variables, Table S2; $MBVIOP_{OnPLS}$ and $VIP_{OPLS-DA}$ values for Block 3 (Metabolomics) variables, Table S3; $MBVIOP_{OnPLS}$ and $VIP_{OPLS-DA}$ values for Block 4 (Fatty Acids) variables, Table S4; $MBVIOP_{OnPLS}$ and $VIP_{OPLS-DA}$ values for Block 5 (Oxylipins) variables, Table S5; $MBVIOP_{OnPLS}$ and $VIP_{OPLS-DA}$ values for Block 6 (Clinical) variables, Table S6; Rho values for correlations between variables presented in Figure 5A (Component 1), Table S7; *P* values for correlations between variables presented in Figure 5A (Component 1), Table S8; Rho values for correlations between variables presented in Figure 5B (Component 2), Table S9; *P* values for correlations between variables presented in Figure 5B (Component 2), Table S10 (XLSX)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: craig.wheelock@ki.se (C.E.W.)

*E-mail: stacey.n.reinke@ecu.edu.au (S.N.R.)

ORCID

Johan Trygg: 0000-0003-3799-6094

Craig E. Wheelock: 0000-0002-8113-0653

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors wish to thank Rickard Sjögren for providing the script to generate the metadata correlation plot. S.N.R. was supported by a Canadian Institutes of Health Research (CIHR) Fellowship (MFE-135481). T.S.C.H. was supported by Wellcome Trust Research Fellowships (088365/z/09/z and 104553/z/14/z), by the Academy of Medical Sciences, and by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). A.S. was supported by the

DOI: 10.1021/acs.analchem.8b03205
Anal. Chem. XXXX, XXX, XXX−XXX

Faculty of Medicine, University of Southampton, UK. B.G.P. was supported by MKS Instruments AB, by IDS/KBC of Umeå University (Sweden) for 2016−2017, and by an ERCIM "Alain Bensoussan" Fellowship Programme at the Department of Engineering Cybernetics (ITK) of the Norwegian University of Science and Technology (Norway) for 2017−2018. C.E.W. was supported by the Swedish Heart Lung Foundation (HLF 20170603). We acknowledge the support of the Swedish Heart Lung Foundation (HLF 20170734), the Swedish Research Council (2016-02798), the Karolinska Institutet, and the ChAMP (Centre for Allergy Research Highlights Asthma Markers of Phenotype) consortium, which is funded by the Swedish Foundation for Strategic Research, the Karolinska Institutet, AstraZeneca & Science for Life Laboratory Joint Research Collaboration, and the Vårdal Foundation.

REFERENCES

(1) Ideker, T.; Galitski, T.; Hood, L. *Annu. Rev. Genomics Hum. Genet.* **2001**, *2*, 343−72.
(2) Kell, D. B.; Oliver, S. G. *BioEssays* **2004**, *26* (1), 99−105.
(3) Brown, M.; Dunn, W. B.; Ellis, D. I.; Goodacre, R.; Handl, J.; Knowles, J. D.; O'Hagan, S.; Spasić, I.; Kell, D. B. *Metabolomics* **2005**, *1* (1), 39−51.
(4) Gromski, P. S.; Muhamadali, H.; Ellis, D. I.; Xu, Y.; Correa, E.; Turner, M. L.; Goodacre, R. *Anal. Chim. Acta* **2015**, *879*, 10−23.
(5) Wheelock, A. M.; Wheelock, C. E. *Mol. BioSyst.* **2013**, *9* (11), 2589−96.
(6) Nørgaard, L.; Bro, R.; Westad, F.; Engelsen, S. B. *J. Chemom.* **2006**, *20* (8−10), 425−435.
(7) Broadhurst, D. I.; Kell, D. B. *Metabolomics* **2007**, *2* (4), 171−196.
(8) Krzanowski, W. J. *Principles of Multivariate Analysis: A User's Perspective*; Clarendon Press, 1988.
(9) Wold, S.; Sjöström, M.; Eriksson, L. *Chemom. Intell. Lab. Syst.* **2001**, *58* (2), 109−130.
(10) Hastie, T.; Tibshirani, T.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer-Verlag: New York, 2009; p 745.
(11) Breiman, L. *Mach Learn* **2001**, *45* (1), 5−32.
(12) Cortes, C.; Vapnik, V. *Mach Learn* **1995**, *20* (3), 273−297.
(13) Bylesjö, M.; Rantalainen, M.; Cloarec, O.; Nicholson, J. K.; Holmes, E.; Trygg, J. *J. Chemom.* **2006**, *20* (8−10), 341−351.
(14) Trygg, J.; Wold, S. *J. Chemom.* **2002**, *16* (3), 119−128.
(15) Madsen, R.; Lundstedt, T.; Trygg, J. *Anal. Chim. Acta* **2010**, *659* (1−2), 23−33.
(16) Xia, J.; Broadhurst, D. I.; Wilson, M.; Wishart, D. S. *Metabolomics* **2013**, *9* (2), 280−299.
(17) Wold, S.; Johansson, E.; Cocchi, M. PLS Partial Least Squares Projections to Latent Structures. In *3D QSAR in Drug Design: Theory, Methods, and Applications*; Kubinyi, H., Ed.; Springer, 1993; pp 523−550.
(18) Galindo-Prieto, B.; Eriksson, L.; Trygg, J. *J. Chemom.* **2014**, *28* (8), 623−632.
(19) Höskuldsson, A.; Svinning, K. *J. Chemom.* **2006**, *20* (8−10), 376−385.
(20) Cavill, R.; Jennen, D.; Kleinjans, J.; Briede, J. J. *Briefings Bioinf.* **2016**, *17* (5), 891−901.
(21) Van Loan, C. F. *SIAM J. Numer Anal* **1976**, *13* (1), 76−83.
(22) Van Deun, K.; Van Mechelen, I.; Thorrez, L.; Schouteden, M.; De Moor, B.; van der Werf, M. J.; De Lathauwer, L.; Smilde, A. K.; Kiers, H. A. L. *PLoS One* **2012**, *7* (5), e37840.
(23) Lock, E. F.; Hoadley, K. A.; Marron, J. S.; Nobel, A. B. *Ann. Appl. Stat* **2013**, *7* (1), 523.
(24) Löfstedt, T.; Trygg, J. *J. Chemom.* **2011**, *25* (8), 441−455.
(25) Smilde, A. K.; Westerhuis, J. A.; de Jong, S. *J. Chemom.* **2003**, *17* (6), 323−337.
(26) Li, C. X.; Wheelock, C. E.; Skold, C. M.; Wheelock, A. M. *Eur. Respir. J.* **2018**, *51* (5), 1701930.
(27) Löfstedt, T.; Hoffman, D.; Trygg, J. *Anal. Chim. Acta* **2013**, *791*, 13−24.
(28) Galindo-Prieto, B. Novel variable influence on projection (VIP) methods in OPLS, O2PLS, and OnPLS models for single-and multi-block variable selection: VIP_{OPLS}, VIP_{O2PLS}, and MB-VIOP methods. Doctoral Dissertation, Umeå University, Umeå, Sweden, 2017.
(29) Hinks, T. S.; Zhou, X.; Staples, K. J.; Dimitrov, B. D.; Manta, A.; Petrossian, T.; Lum, P. Y.; Smith, C. G.; Ward, J. A.; Howarth, P. H.; Walls, A. F.; Gadola, S. D.; Djukanovic, R. *J. Allergy Clin. Immunol.* **2015**, *136* (2), 323−33.
(30) Reinke, S. N.; Gallart-Ayala, H.; Gomez, C.; Checa, A.; Fauland, A.; Naz, S.; Kamleh, M. A.; Djukanovic, R.; Hinks, T. S.; Wheelock, C. E. *Eur. Respir. J.* **2017**, *49* (3), 1601740.
(31) Vijayanand, P.; Seumois, G.; Pickard, C.; Powell, R. M.; Angco, G.; Sammut, D.; Gadola, S. D.; Friedmann, P. S.; Djukanovic, R. *N. Engl. J. Med.* **2007**, *356* (14), 1410−22.
(32) Diez, D.; Wheelock, A. M.; Goto, S.; Haeggstrom, J. Z.; Paulsson-Berne, G.; Hansson, G. K.; Hedin, U.; Gabrielsen, A.; Wheelock, C. E. *Mol. BioSyst.* **2010**, *6* (2), 289−304.
(33) Skotare, T.; Sjögren, R.; Surowiec, I.; Nilsson, D.; Trygg, J. *J. Chemom.* **2018**, e3071.
(34) Wold, S.; Kettaneh, N.; Tjessem, K. *J. Chemom.* **1996**, *10* (5−6), 463−482.
(35) Galindo-Prieto, B.; Trygg, J.; Geladi, P. *Chemom. Intell. Lab. Syst.* **2017**, *160*, 110−124.
(36) Holten, D. *IEEE Trans Vis Comp Graph* **2006**, *12* (5), 741−748.
(37) Hunter, J. D. *Comput. Sci. Eng.* **2007**, *9* (3), 90−95.
(38) Wang, D.; Bodovitz, S. *Trends Biotechnol.* **2010**, *28* (6), 281−90.
(39) Holgate, S. T.; Wenzel, S.; Postma, D. S.; Weiss, S. T.; Renz, H.; Sly, P. D. *Nat. Rev. Dis Primers* **2015**, *1*, 15025.
(40) Tan, L. J.; Wang, Z. E.; Wu, K. H.; Chen, X. D.; Zhu, H.; Lu, S.; Tian, Q.; Liu, X. G.; Papasian, C. J.; Deng, H. W. *J. Clin. Endocrinol. Metab.* **2015**, *100* (11), E1457−66.
(41) Wong, C. A.; Walsh, L. J.; Smith, C. J.; Wisniewski, A. F.; Lewis, S. A.; Hubbard, R.; Cawte, S.; Green, D. J.; Pringle, M.; Tattersfield, A. E. *Lancet* **2000**, *355* (9213), 1399−1403.
(42) McNamara, P.; Seo, S. B.; Rudic, R. D.; Sehgal, A.; Chakravarti, D.; FitzGerald, G. A. *Cell* **2001**, *105* (7), 877−89.
(43) Jang, Y. S.; Kang, Y. J.; Kim, T. J.; Bae, K. *Mol. Biol. Rep.* **2012**, *39* (4), 4215−21.
(44) Gooley, J. J.; Chua, E. C. *J. Genet. Genomics* **2014**, *41* (5), 231−50.
(45) Kenfield, M.; Yu, H.; Ehlers, A.; Xie, W.; Gunsten, S.; Agapov, E.; Horani, A.; Holtzman, M. J.; Brody, S. L.; Haspel, J. *Am J Respir Crit Care Med* **2017**, *195*, A5210.

DOI: 10.1021/acs.analchem.8b03205
Anal. Chem. XXXX, XXX, XXX−XXX