Journal of proteome research

Article

pubs.acs.org/jpr

# moCluster: Identifying Joint Patterns Across Multiple Omics Data Sets

Chen Meng,[†] Dominic Helm,[†] Martin Frejno,[†,‡] and Bernhard Kuster*[,†,§,∥]

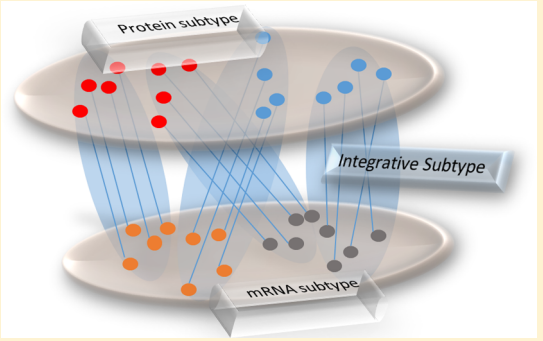
[†]Chair of Proteomics and Bioanalytics, [§]Bavarian Biomolecular Mass Spectrometry Center, Technische Universität München, Freising 85354, Germany

[‡]Department of Oncology, University of Oxford, Oxford OX3 7DQ, United Kingdom

[∥]Center for Integrated Protein Science Munich (CIPSM), Emil-Erlenmeyer-Forum 5, Freising 85354, Germany

Supporting Information

**ABSTRACT:** Increasingly, multiple omics approaches are being applied to understand the complexity of biological systems. Yet, computational approaches that enable the efficient integration of such data are not well developed. Here, we describe a novel algorithm, termed moCluster, which discovers joint patterns among multiple omics data. The method first employs a multiblock multivariate analysis to define a set of latent variables representing joint patterns across input data sets, which is further passed to an ordinary clustering algorithm in order to discover joint clusters. Using simulated data, we show that moCluster's performance is not compromised by issues present in iCluster/iCluster + (notably, the nondeterministic solution) and that it operates 100× to 1000× faster than iCluster/iCluster+. We used moCluster to cluster proteomic and transcriptomic data from the NCI-60 cell line panel. The resulting cluster model revealed different phenotypes across cellular subtypes, such as doubling time and drug response. Applying moCluster to methylation, mRNA, and protein data from a large study on colorectal cancer patients identified four molecular subtypes, including one characterized by microsatellite instability and high expression of genes/proteins involved in immunity, such as PDL1, a target of multiple drugs currently in development. The other three subtypes have not been discovered before using single data sets, which clearly illustrates the molecular complexity of oncogenesis and the need for holistic, multidata analysis strategies.




**KEYWORDS:** *Multiple omics data, clustering, cancer, stratification, data analysis*

## INTRODUCTION

Cancer is a heterogeneous disease. Even when it originates from the same tissue, the underlying molecular mechanisms can vary dramatically between patients. Therefore, every patient should receive a personalized treatment strategy according to a (molecular) stratification scheme. Broad profiling of genetic mutations, mRNAs, proteins, and other biological molecules provides valuable sources for the stratification of cancer patients into such molecular subgroups. However, due to the complexity of the underlying biology, a study focusing only on one of the different molecular levels (e.g., genome, transcriptome, or proteome) may fail to reveal important factors contributing to oncogenesis. Conversely, an integrated analysis using multiple levels of information may provide a much more detailed picture not readily available from a single data set.[1] Therefore, an increasing number of projects such as The Cancer Genome Atlas (TCGA), the Cancer Cell Line Encyclopedia (CCLE), and the Encyclopedia of DNA Elements (ENCODE) aims to systematically measure multiple levels of omics data from a large number of samples in an attempt to derive a comprehensive understanding of the mechanisms underlying oncogenesis. Perhaps not surprisingly, clustering based on different types of data almost always results in different subtype models. Hence, there is an increasing need for methods capable of integrative clustering of multiple data sets.

Recently, researchers have used the clustering of clusters (COCL) algorithm and identified 13 integrative subtypes (including copy number variation, DNA methylation, mRNA, miRNA, and protein expression data) from thousands of cancer samples originating from 12 different tumor sites.[2] The COCL algorithm is a simple method involving two steps of clustering. First, a clustering algorithm is performed on each individual omics data set. The clustering results are represented by dummy matrices that contain binary vectors indicating the cluster assignments for each subtype. Next, the dummy matrices for all omics data are concatenated and passed to a clustering method (such as consensus clustering[3]) to identify

**Special Issue:** Large-Scale Computational Mass Spectrometry and Multi-Omics

**Received:** September 4, 2015
**Published:** December 14, 2015






DOI: 10.1021/acs.jproteome.5b00824
J. Proteome Res. 2016, 15, 755−765

the joint pattern of multiple omics data. However, this study failed to detect the common pathways or mechanisms shared by cancers from different origins. Traditional methods for analyzing two data sets rely on the analysis of a correlation matrix.[4] Shen et al. criticized correlation matrix-based methods as aiming to identify the correlated pattern, which is insufficient for the identification of unique or complementary patterns in each data set.[5] Therefore, the iCluster algorithm was proposed to address this issue.[5] In the iCluster framework, multiple high-dimensional data sets are represented by a low number of common variables, also called latent variables. The latent variables may account for distinct molecular subtype-related biological molecules in each data set. Therefore, a clustering of the latent variables is able to represent the integrated patterns of multiple data sets.[5] The method has already been successfully applied in several studies. For example, an integrative analysis of copy number and gene expression data of 2000 breast cancer tissues resulted in a novel subtype model with distinct clinical relevance.[6] Recently, an extension of the iCluster algorithm, called iCluster+, was developed to cluster a more diverse range of data types.[7] In iCluster+, different models are used to account for diverse data types, for example, logistic regression for binary variables (mutation data), the multilogit regression model for multicategory variables (copy number variation), or Poisson regression for count variables (sequencing data).[7] Despite their widespread application, there are limitations for these methods. For example, they use an iterative expectation-maximization algorithm, which does not necessarily converge to a deterministic or optimal solution. In addition, these algorithms are computational intensive, which is a particular limitation to nondeterministic algorithms since the only remedy for nondeterministic results is to run the algorithm multiple times and select a consistent output.

In this study, we introduce a new method, termed moCluster, which is based on a multitable multivariate analysis. We evaluated the iCluster/iCluster+ algorithms and compared them to our novel algorithm using simulated data sets. We demonstrate that moCluster outperforms iCluster/iCluster+ since moCluster defines a subspace that distinguishes subtypes more clearly and always converges to a deterministic solution. At the same time, moCluster is 100× to 1000× faster than the other two algorithms. We also applied our method to the transcriptomic and proteomic data sets of the NCI-60 cell line panel and to methylation, transcriptomics, and proteomics data from the TCGA colorectal cancer study. The latter analysis resulted in a four-subtype model, of which one was previously known, whereas the others could be discovered only based on the combined data. Therefore, this work not only provides a novel method for the integration of multiple levels of omics data but also forms the starting point for future research on newly discovered molecular subtypes.

## METHODS

### moCluster Approach

The first step of the moCluster algorithm is defining a joint latent variable (JLV) using the modified consensus PCA (CPCA), which is calculated using a multiple block extension of the NIPALS algorithm. We describe the algorithm using notations by Westerhuis et al.[8] The input of the algorithm is a set of matrices ($\boldsymbol{X}_1$, $\boldsymbol{X}_2$, ..., $\boldsymbol{X}_k$, ..., $\boldsymbol{X}_K$), where rows are observations and columns are features (genes/proteins). In the algorithm, $\boldsymbol{t}$ is the JLV and $\boldsymbol{p}_k$ and $\boldsymbol{t}_k$ are the feature coefficients vector (loading) and the block latent variables (BLVs) for matrix $k$, respectively.

*Transform, center and scale*
*For each latent variable:*

1. *Randomly choose start* $\boldsymbol{t}$
2. *loop until convergence of* $\boldsymbol{t}$ *:*
   - 2.1 $\boldsymbol{p}_k = \boldsymbol{X}_k^T \cdot \boldsymbol{t} \,/ \boldsymbol{t}^T\boldsymbol{t}$
   - 2.2 *normalize* $\mathbf{p}_k$ *to* $\|\boldsymbol{p}_k\| = 1$
   - 2.3 $\boldsymbol{p}_k = \mathrm{soft}(\boldsymbol{p}_k)$ # soft-thresholding operator, introduce sparsity
   - 2.4 $\boldsymbol{t}_k = \boldsymbol{X}_k \cdot \boldsymbol{p}_k$
   - 2.5 $\boldsymbol{T} = [\boldsymbol{t}_1, \ldots, \boldsymbol{t}_K]$
   - 2.6 $\boldsymbol{w}_T = \boldsymbol{T} \cdot \boldsymbol{t} \,/ \boldsymbol{t}^T \cdot \boldsymbol{t}$ # $\mathbf{w}_T$ is the linear combination coefficients of BLV to construct the JLV
   - 2.7 *normalize* $\mathbf{w}_T$ *to* $\|\boldsymbol{w}_T\| = 1$
   - 2.8 $\boldsymbol{t} = \boldsymbol{T} \cdot \boldsymbol{w}_T$

   *end*
3. $\boldsymbol{p}_k = \boldsymbol{X}_k^T \boldsymbol{t} \,/ \boldsymbol{t}^T \cdot \boldsymbol{t}$ # final coefficients for features in matrix $k$
4. $\boldsymbol{X}_k = \boldsymbol{X}_k - \boldsymbol{t} \cdot \boldsymbol{p}_k^T$ # deflation by JLV

*end*

In addition to the centering and scaling of features in each data set, the integrative analysis of multiple data sets requires a normalization on data set level because the one with more features (columns) often has more overall variance. In this study, we solved this issue by weighting each data matrix by the reverse of its first eigenvalue, allowing different matrices to contribute comparable variance to the first (few) JLV(s). After proper normalization of the data sets, the algorithm calculates JLVs in a sequential manner. First, a random JLV $\boldsymbol{t}$ is initialized (step 1), which is further iteratively updated until convergence is reached (steps 2.1−2.8). In each iteration, the coefficient vector $\boldsymbol{p}_k$ (loading) of each individual matrix $\boldsymbol{X}_k$ is derived by regressing $\boldsymbol{X}_k$ to the initialized $\boldsymbol{t}$ (step 2.1). Then, the BLV for data set $k$ is calculated by regressing $\boldsymbol{X}_k$ to the coefficient vector $\boldsymbol{p}_k$ (steps 2.2−2.4). We modified the CPCA approach for feature selection by introducing a soft-thresholding operator in this process (step 2.3). The soft(·) is defined as $\mathrm{soft}(\boldsymbol{x}, a) = \mathrm{sin}(\boldsymbol{x})(|\boldsymbol{x}| - a)_+$. By tuning parameter $a$, we explicitly control the number of nonzero coefficients in $\boldsymbol{p}_k$. Steps 2.5−2.8 show that JLV $\boldsymbol{t}$ is updated according to the weights computed from the BLVs $\boldsymbol{t}_k$. Steps 2.1−2.8 are iterated until convergence is reached. Hanafi et al. have previously shown that this algorithm is convergent.[25] The higher order JLVs are based on the residual matrices, which are calculated in step 4. The different ways of calculating residual matrices underscore the difference among CPCA, GCCA, and MCIA: GCCA deflates matrices with respect to BLVs ($\boldsymbol{t}_k$; i.e., step 4 should be $\boldsymbol{X}_k = \boldsymbol{X}_k - \boldsymbol{t}_k \cdot \boldsymbol{p}_k^T$), which do not need to be directly related to the variation of the global variation pattern and MCIA deflates matrices with respect to coefficient vectors ($\boldsymbol{p}_k$; step 4 should be $\boldsymbol{X}_k = \boldsymbol{X}_k - \boldsymbol{X}_k \cdot \boldsymbol{p}_k \cdot \boldsymbol{p}_k^T$).[9] The advantage of deflation matrices with respect to JLV $\boldsymbol{t}$, as in CPCA, is that this strategy guarantees the orthogonality of JLVs. At the same time, the coefficient vectors can incorporate a sparse operator.

Next, the JLVs can be clustered by conventional clustering algorithms, such as $K$-means or consensus clustering.[3] In this study, we used the hierarchical clustering algorithm because of its simplicity (the Euclidean distance measurement and Ward linkage method[26]). The optimal number of clusters was evaluated by the gap-statistic.[12]

**Software Information.** The method is implemented in the Bioconductor package "mogsa". The package requires R, version 3.2.0 or newer, and is operating system independent. It can be download from https://www.bioconductor.org/


DOI: 10.1021/acs.jproteome.5b00824
J. Proteome Res. 2016, 15, 755−765

packages/release/bioc/html/mogsa.html. The package also includes sample data sets, a user help manual, and a vignette. The vignette (https://www.bioconductor.org/packages/release/bioc/vignettes/mogsa/inst/doc/moCluster.pdf) describes a typical workflow using the sample data sets. The user manual that describes the usage of functions can be downloaded from https://www.bioconductor.org/packages/release/bioc/manuals/mogsa/man/mogsa.pdf. moCluster is under a GPL-2 license.

### iCluster

The iCluster/iCluster+ algorithms were called from the CRAN package "iCluster" (version 2.1.0) and the Bioconductor R package "iClusterPlus" (version 1.2.0), respectively. The iCluster function in the iCluster package implemented the method described in Shen et al. (2009) and Mo et al. (2013).[5,7]

## Data Acquisition and Processing

**NCI-60 Data.** The NCI-60 drug sensitivity data (DTP NCI-60 *z*-scores) and mRNA data (average *z*-score from five microarray platforms) were downloaded from CELLMINER (download date: 2014-06-23).[11] The proteomics data were downloaded from the supplementary table of Moghaddas Gholami et al. (2013).[10] The proteome data were quantified and normalized using the iBAQ method.[27] Because the proteomics data for the melanoma cell line MDAN and the mRNA data for the CNS cell line SNB19 were not available, the two cell lines were excluded from the analysis, resulting in 58 cell lines. Missing values in the transcriptomic and proteomic data were replaced with zero. The mRNA expression data were filtered based on the standard deviation of a gene across the 58 cell lines; genes with standard deviation greater than 0.8 were retained. In the proteomics data, proteins with a total intensity across the panel greater than 10 were retained. The filtering of the original data, which contain 26 065 genes (transcriptome data) and 8113 proteins (proteome data), resulted in 11 826 and 8069 genes and proteins, respectively. Then, the iBAQ values were transformed by $x_i = \log_{10}(\text{iBAQ}_i + 1)$.

**TCGA Colorectal Cancer Data.** To evaluate the integrative subtypes of colorectal cancers, the mRNA expression (RPKM) matrix and methylation matrix were downloaded from the cBio cancer genomics portal.[28] The spectral count proteomics data and related subtypes, oncogene mutation, and clinical information were downloaded from the Supporting Information of Zhang et al. (2014).[16b] Patients with all three types of data (mRNA, methylation, and proteomics) were retained in the analysis, resulting in 83 patients. The association between each integrated subtype and other factors was evaluated by two-sided Fisher's exact test.

The methylation level was represented by the beta-value. Beta-values were transformed to *M*-values as $M_i = \log_2(\text{beta}_i/(1 - \text{beta}_i))$.[29] The RPKM values were transformed by $x_i = \log_{10}(\text{RPKM}_i + 1)$. Then, all data underwent an unspecific filtering. For each methylation point in the methylation data, the sum of *M*-values across the 83 patients was calculated. We filtered out methylation sites with a sum of *M*-values lower than −300. In addition, due to the distinct methylation status of the X chromosome between genders, we removed all methylation sites located on the X chromosome. For the mRNA expression data, genes with median absolution deviation (MAD) greater than 0.1 were retained. Due to the presence of missing values in the proteomics data, we selected proteins with less than 60 missing values. Filtering of 27 578, 20 533, and 7211 unique features in the original methylation, mRNA, and proteomics data sets resulted in 11 282, 12 503, and 5708 sites/genes/proteins, respectively.

### Simulation of simData1

In order to simulate data that mimic the true variance level of real data and at the same time have a clearly defined cluster pattern, we used the following procedure.

We downloaded the TCGA bladder cancer transcriptomic and CNV data using TCGA assembler (date: 26/09/2014). The transcriptome data (RNaseqV2) were quantified by the MapSplice and RSEM approaches and were subsequently logarithm transformed. The gene level CNV is estimated by the mean copy number of the genomic region corresponding to a gene (retrieved by TCGA assembler directly[30]). We randomly selected 3000 genes and 240 matched patients in the two data sets, referred to as $\boldsymbol{X}_{\text{cnv}}$ and $\boldsymbol{X}_{\text{mrna}}$, where rows are patients and columns are genes. Then, the two matrices underwent singular value decomposition (SVD)

$$\boldsymbol{X}_{\text{cnv}} = \boldsymbol{U}_{\text{cnv}}\boldsymbol{D}_{\text{cnv}}\boldsymbol{V}_{\text{cnv}}^T \text{ and } \boldsymbol{X}_{\text{mrna}} = \boldsymbol{U}_{\text{mrna}}\boldsymbol{D}_{\text{mrna}}\boldsymbol{V}_{\text{mrna}}^T$$

$\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices known as the left and right singular matrix; their columns are singular vectors. $\boldsymbol{D}$ is the diagonal matrix, where the diagonal elements represent the significance of the corresponding singular vectors.

We planned to clearly define two subtypes in each of the data sets, which should be described by the first left singular vector. In order to reflect the exceptional importance of the first left singular vectors and avoid the interference of other singular vectors, we modified the diagonal elements of $\boldsymbol{D}$: we retained the first diagonal elements in $\boldsymbol{D}$, and the other diagonal elements were replaced by their means, that is, $\hat{d}_i = \begin{cases} d_i & \textbf{if} \quad i = 1 \\ \text{mean}(d_{i\neq 1}) & \textbf{if} \quad i \neq 1 \end{cases}$ for the *i*th diagonal elements $d_i$. Therefore, the pattern defined by the first singular vector represents the true signal, and the remaining singular vectors are noise. The modified diagonal singular value matrices were denoted $\hat{\boldsymbol{D}}_{\text{cnv}}$ and $\hat{\boldsymbol{D}}_{\text{mrna}}$

However, the first singular vectors in $\boldsymbol{U}_{\text{cnv}}$ and $\boldsymbol{U}_{\text{mrna}}$ do not have a clear cluster structure. Therefore, we needed to simulate $\boldsymbol{U}_{\text{cnv}}$ and $\boldsymbol{U}_{\text{mrna}}$ as well. To do so, we first defined two matrices, $\boldsymbol{X}_{\text{sim1}}$ and $\boldsymbol{X}_{\text{sim2}}$, which have the same dimensions as $\boldsymbol{X}_{\text{cnv}}$ and $\boldsymbol{X}_{\text{mrna}}$, by a simple linear additive model

$$x_{ij} = \tilde{x}_{ij} + \varepsilon_{ij}$$

where $\varepsilon_{ij} \sim \mathcal{N}(0,1)$ represents random noise, $\tilde{x}_{ij} \sim \mathcal{N}(0,\text{SD}_{\text{signal}})$ is the pseudo gene expression level, and $\tilde{x}_{ij_1} = \tilde{x}_{ij_2}$ if $j_1$ and $j_2$ belong to the same cluster. In $\boldsymbol{X}_{\text{sim1}}$, the two clusters are samples 1−120 and 121−240, whereas in $\boldsymbol{X}_{\text{sim2}}$, the two clusters are samples 1−60/121−180 and 61−120/181−240. Therefore, the two data sets define four subtypes as samples 1−60, 61−120, 121−180, and 181−240. In order to create orthogonal matrices to replace $\boldsymbol{U}_{\text{cnv}}$ and $\boldsymbol{U}_{\text{mrna}}$, we calculated the SVD of $\boldsymbol{X}_{\text{sim1}}$ and $\boldsymbol{X}_{\text{sim2}}$

$$\boldsymbol{X}_{\text{sim1}} = \boldsymbol{U}_{\text{sim1}}\boldsymbol{D}_{\text{sim1}}\boldsymbol{V}_{\text{sim1}}^T \text{ and } \boldsymbol{X}_{\text{sim2}} = \boldsymbol{U}_{\text{sim2}}\boldsymbol{D}_{\text{sim2}}\boldsymbol{V}_{\text{sim2}}^T$$

Finally, the simulated CNV and mRNA data were generated by

$$\boldsymbol{X}_{\text{simCNV}} = \boldsymbol{U}_{\text{sim1}}\hat{\boldsymbol{D}}_{\text{cnv}}\boldsymbol{V}_{\text{cnv}}^T \text{ and } \boldsymbol{X}_{\text{simMRNA}} = \boldsymbol{U}_{\text{sim2}}\hat{\boldsymbol{D}}_{\text{mrna}}\boldsymbol{V}_{\text{mrna}}^T$$

Therefore, the first left singular vectors in $\boldsymbol{U}_{\text{sim1}}$ and $\boldsymbol{U}_{\text{sim2}}$ capture the two-cluster structure in each data set; the four-



DOI: 10.1021/acs.jproteome.5b00824
J. Proteome Res. 2016, 15, 755−765

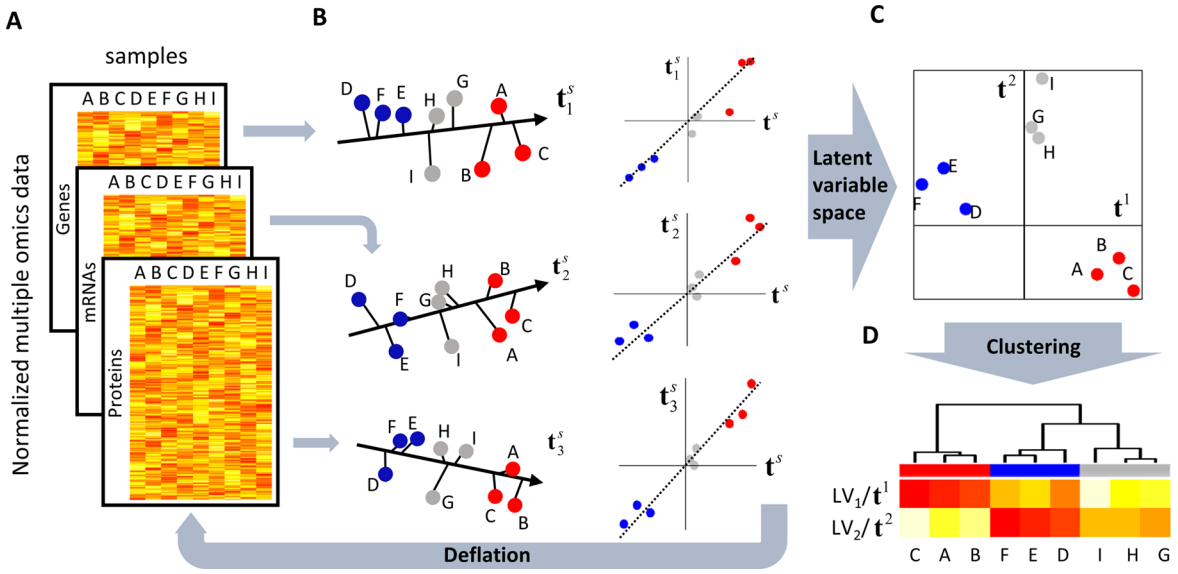



**Figure 1.** Schematic view of the moCluster algorithm. The input of the CPCA algorithm is a set of matrices that have matched columns, such as genomic, transcriptomics, and proteomics data describing the same patient cohort (A). (B) Consensus PCA (CPCA) approach. For a latent variable *s*, CPCA uses a linear combination of original features to define a set of block latent variables (BLVs; $t_1^s$, $t_2^s$, and $t_3^s$) and a joint latent variable (JLV; $t^s$) to maximize the summed correlation between each BLV and the JLV. To derive the higher order solution, the variance accounted for by the JLV is removed from the matrix (deflation step) and the same process reiterated until all of the latent variables are defined. (C) Scatter plot showing the space defined by two latent variables. (D) Application of a clustering algorithm on the latent variables to find the joint patterns across data sets. Different colored points indicate different cluster members.

cluster structure defined by $X_{\text{simCNV}}$ and $X_{\text{simMRNA}}$ could be represented by exactly two latent variables. To simulate different signal-to-noise ratios, we defined $SD_{\text{signal}}$ = 1, 0.5, and 0.2 for high, medium, and low signal-to-noise ratios, respectively.

In addition, to simulate data with sparsity, 1000 genes were randomly selected in the first column of $V_{\text{cnv}}{}^T$ and $V_{\text{mrna}}{}^T$ as nonsparse genes. All other values were set to 0. Then, the vectors were rescaled so that the sum of square of all values equals one.

## RESULTS AND DISCUSSION

### moCluster Approach

The steps taken in the moCluster algorithm can be summarized as follows (Figure 1):

(1) Use sparse consensus PCA to find latent variables.
(2) Use permutation and elbow test to determine the number of latent variables.
(3) Cluster latent variables (using, e.g., hierarchical or *K*-means clustering).
(4) Select the best subtype model.

When multiple omics data sets are available for the same set of observations, the data can be represented by a set of matrices ($X_1$, $X_2$, ..., $X_K$). The rows of the matrices refer to the same set of observations (e.g., samples, cell lines, or patients), whereas the columns refer to different features such as genes, mRNAs, or proteins (Figure 1A). The core idea of moCluster is similar to iCluster. Both use linear combinations of original features (variables) to define a set of JLVs, which represent the most important patterns as defined by multiple omics data.[5] This can be expressed as

$$\begin{aligned} X_1 &= TP_1{}^T + E_1 \\ X_2 &= TP_2{}^T + E_2 \\ &\vdots \\ X_K &= TP_K{}^T + E_K \end{aligned} \quad (1)$$

where $T = |t^1, ..., t^s, ..., t^S|$ is a matrix that comprises the JLVs $t^s$ in columns and ($P_1$, $P_2$, ..., $P_K$) are the matrices of coefficients for features in each data sets (also known as loading matrices). In contrast to iCluster, which uses an expectation-maximization algorithm to optimize a log-likelihood function derived from this model, the moCluster algorithm employs the consensus PCA (CPCA[8]) approach to estimate the latent variables. In this study, we modified the CPCA algorithm to introduce sparsity in feature coefficient vectors (columns of $P_k$; see Methods) in order to facilitate the biological interpretation of clustering results. For a single matrix, principal component analysis (PCA) models the high dimensional data with a lower number of variables (principal components, PC). The PCs discovered by PCA are the optimal representation for a single matrix and are widely used in conjunction with clustering or regression methods, such as spectral clustering. CPCA is an extension of PCA for the analysis of multiple matrices. In order to calculate the first latent variable, CPCA finds a set of suboptimal latent variables for each individual matrix $t_k^1$ (denoted block latent variable; $t_1^s$, $t_2^s$, $t_3^s$ in Figure 1B) by regressing individual matrices to the their respective feature coefficient vectors (see Methods). The BLVs are suboptimal because they are not necessarily the best representation of the matrix itself. Instead, CPCA aims at finding a JLV ($t^s$ in Figure 1B) via a linear combination of BLVs so that the summed covariance between BLVs and JLV, i.e., $\sum_{k=1}^{K} \text{cov}^2(t_k^1, t^1)$, is maximized.[9] As a result, the JLVs represent the joint patterns of multiple data sets. To calculate a subsequent (higher order) JLV, CPCA first computes the residual matrices from each original matrix by


DOI: 10.1021/acs.jproteome.5b00824
J. Proteome Res. 2016, 15, 755−765

removing the variance that accounted for the previously calculated JLV. This procedure is called deflation. Next, the JLV and feature coefficients are calculated from the residual matrices using the same procedure as before. These processes are repeated until the desired number of JLVs is calculated. Figure 1C shows an example of a two-dimensional JLV space, and the detailed algorithm is described in the Methods section.

Like PCA, the importance of JLVs can be measured by their pseudoeigenvalues (explained variance), which monotonically decrease by definition. Frequently, the most important JLVs are the ones before or at the "elbow" point in a scree plot, where the slope of pseudoeigenvalues decreases from steep to flat. We used this method to determine the number of JLVs that should be included in the cluster analysis (see also the application studies below). In addition, a permutation test can be used to evaluate the concordant or divergent structures between data sets. To do so, moCluster shuffles all of the observations/samples (rows) in each of the data sets, leading to the random matching of observations in different data sets. The permuted data sets are then passed to the CPCA. Bootstrapping may also be applied in this step. However, we used random permutation in this study because bootstrapping may inflate the compactness (within-cluster vs between-cluster variability) of the clustering results.[3] An empirical confidence interval for each pseudoeigenvalue is derived by repeating the permutation multiple times. The permutation analysis provides a further reference for choosing the JLV. Eigenvalues significantly higher than the permutation eigenvalues represent the concordant structures across the data sets, whereas including extra JLVs enables the detection of divergent structures in multiple data sets. In practice, an elbow test in combination with the permutation test is used to determine the necessary number of JLVs.

With the principle of parsimony in mind, we modified the CPCA algorithm by introducing a soft-thresholding operator to ensure sparsity of the feature coefficients ($\boldsymbol{P}_k$ in eq 1) for each of the JLVs (see Methods). The sparsity of the coefficient matrix leads to the violation of the maximized covariance criteria (i.e., will explain less variance), but it greatly facilitates the interpretation of the biological meaning of the JLVs. In the last step, an ordinary clustering algorithm is applied on the JLVs (Figure 1D).

A robust estimation of the JLVs is crucial for integrative clustering, and this is the essential difference between the moCluster and iCluster/iCluster+ approaches. Several other generalizations of PCA for multiple-table problems have been proposed and applied to omics data analysis, including generalized canonical correlation analysis (GCCA) and multiple co-inertia analysis (MCIA).[1,9] These methods use different deflation strategies for the calculation of JLVs.[5] The deflation step in GCCA relies on BLVs, leading to nonorthogonal JLVs. This is unfavorable since it indicates that different JLVs can be driven by the same or a subset of correlated features (e.g., genes or proteins). In the MCIA approach, the residual matrices are calculated based on the feature coefficient vectors (columns of $\boldsymbol{P}_k$ in eq 1; see Methods). As a result, a sparse solution of feature coefficients may dramatically influence the JLVs, often leading to sparse and correlated JLVs. Therefore, the CPCA approach is particularly suitable for applications in integrative clustering.

## Comparison of moCluster to iCluster/iCluster+ Using Simulated Data

In order to have exact control over a putative molecular subtype pattern in molecular profiling data of, say, cancer patients, we first used simulated data (referred to as simData1) to compare moCluster to iCluster and iCluster+. Two data sets (Figure 2A), each consisting of 240 samples, were generated to represent two omics data sets (Data 1 and Data 2). Data 1 was further divided into two subclusters: the first subcluster contained samples 1−120, whereas the second cluster consisted of samples 121−240. Data 2 also consists of two subclusters: the first cluster is composed of samples 1−60 and 121−180, whereas the second cluster consists of samples 61−120 and 181−240. Thus, a combination of Data 1 and Data 2 results in four different integrated subtypes, namely, samples 1−60, 61−120, 121−180, and 181−240 (the light blue, orange, blue, and red color bars in Figure 2A). In order to facilitate comparison and visualization, these data sets were simulated in such a way that the four-cluster pattern could be captured within two JLVs (see Methods). In order to determine the sensitivity of each method, the data were simulated three times with varying signal-to-noise ratios (see Methods). We then used the three algorithms to calculate two JLVs. The corresponding results are shown in Figure 2B−D. We observed that the moCluster algorithm can always distinguish the four clusters defined by the two data sets. In line with expectation, as soon as the signal-to-noise ratio becomes lower, the samples are more disperse in

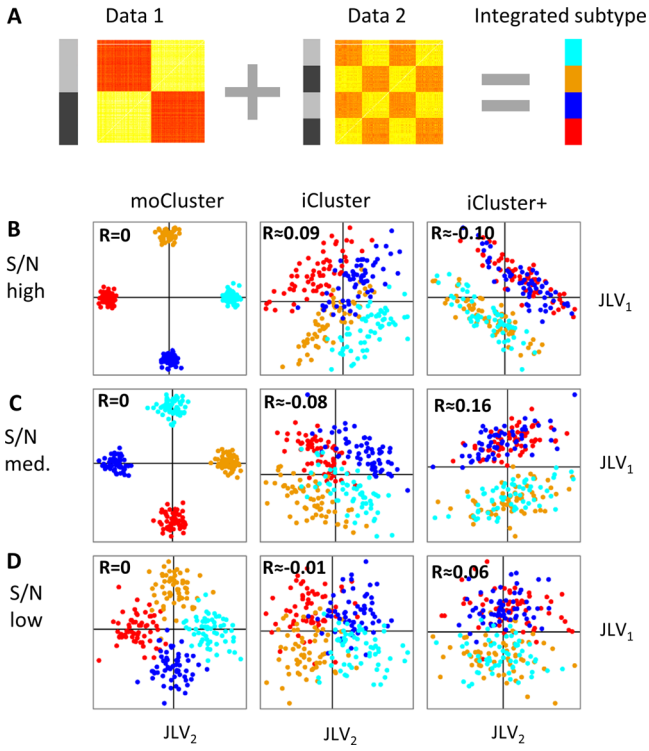



**Figure 2.** Comparison of moCluster to iCluster and iCluster+ using simulated data. (A) Two data sets were simulated, each of them consisting of two clusters. The combination of them resulted in four clusters, as shown by different colors. (B) The space defined by the two latent variables discovered by moCluster, iCluster, and iCluster+ on high signal-to-noise ratio data. (C, D) The same as in (B) but the three algorithms were applied to data with different signal-to-noise ratios (S/N). *R* indicates the correlation coefficient.



DOI: 10.1021/acs.jproteome.5b00824
*J. Proteome Res.* 2016, 15, 755−765

their defined clusters. The iCluster algorithm can also roughly separate the four subtypes of simulated samples; however, it does not benefit from increasing the signal-to-noise ratio. Interestingly, the iCluster+ algorithm was more likely to discover the two subtypes defined by Data 1. Apart from a better discovery of the joint patterns, the moCluster algorithm also runs 100× to 1000× times faster than the other two algorithms (Table 1). The long computation time is particularly

**Table 1. Computation Time Comparison of Different Algorithms**

| | time (s)[a] | | |
|---|---|---|---|
| data description[b] | moCluster | iCluster[c] | iCluster+[c] |
| S/N high | 0.8 | 713.4 | 1042.4 |
| S/N mid | 1.0 | 712.5 | 1016.6 |
| S/N low | 1.0 | 711.9 | 992.7 |

[a]1 core; Intel(R) Core(TM) i5 CPU 650 @ 3.20 GHz 16 GB RAM. [b]Two data sets, each of them has 240 samples and 3000 variables. [c]The number of maximum iterations is 10.

problematic for iCluster+. This is because iCluster+ uses a Monte Carlo Newton−Raphson algorithm, which produces nondeterministic results due to the Monte Carlo sampling procedure (Figure S1).[7] To address this problem, the algorithm should be run multiple times and the best result (i.e., the most consistent) can then be selected according to a defined criterion. However, the long computation time requirement may impair the application of iCluster+ to the analysis of very large omics data sets.

In this study, a maximum of 10 iterations was initially allowed for iCluster/iCluster+, but the algorithms did not necessarily converge after 10 iterations. To evaluate the impact of the allowed iteration number, we increased the maximum number of allowed iterations from 10 to 50 for both methods. The result suggested that iCluster with a maximum 50 iterations resulted in highly correlated latent variables; even more problematic was that the algorithm converged to a solution that distinguished only two subtypes (Figure S2). In fact, although both methods try to define orthogonal latent variables,[5,7] the results can still correlate with each other (Figure 2B). For the iCluster+ algorithm, the increase in the number of iterations did not affect the results at all, whereas the results from iCluster may be strongly influenced by the number of iterations. In practice, a proper number of iterations may need to be determined so that the results represent a good trade-off between processing time and confidence in the results.

To further evaluate the performance of different integrative clustering methods, we applied each algorithm to more realistic data based on the proteomes and transcriptomes of the SNB75 (central nervous system, CNS) and K562 (leukemia) cell lines (referred to as simData2). In this analysis, a random subset of transcriptomic and proteomic data of each cell line (1000 genes or proteins) was selected and duplicated 12 times. The 12 transcriptomic duplicates of SNB75 were divided into two groups (six in each), which were matched to proteomic data from SNB75 and K562, respectively. Similarly, the transcriptome profiling of K562 was matched to proteomic data of both cell lines (Table 2). As a result, we created two 24 × 1000 matrices (gene and protein expression matrices) that, together, represent four clusters. Similar to the simpler simulation data discussed above, the top two JLVs represent this four-cluster structure. We further introduced experimental

**Table 2. Cluster Design of simData2[a]**

| | | proteome (P) | |
|---|---|---|---|
| | | SNB75 | K562 |
| transcriptome (T) | SNB75 | 6 rep. | 6 rep. |
| | K562 | 6 rep. | 6 rep. |

[a]The four subtypes are T/SNB75 + P/SNB75, T/SNB75 + P/K562, T/K562 + P/SNB75, and T/K562 + P/K562.

noise to the matrices by randomly sampling from normal distributions with a mean of zero and different standard deviations, thus allowing for the simulation of different noise levels. For each selected noise level, we applied moCluster, iCluster, and iCluster+ to 100 data sets. The performance of each algorithm was measured by the clustering accuracy, defined as acc = (number of correct cluster assignments/total number of cluster assignments). The results show that moCluster achieved significantly higher accuracy than the iCluster and iCluster+ methods (Wilcoxon rank-sum test; Figure 3). iCluster+ exhibited the least accuracy, probably in part because of the nondeterministic results of this algorithm.

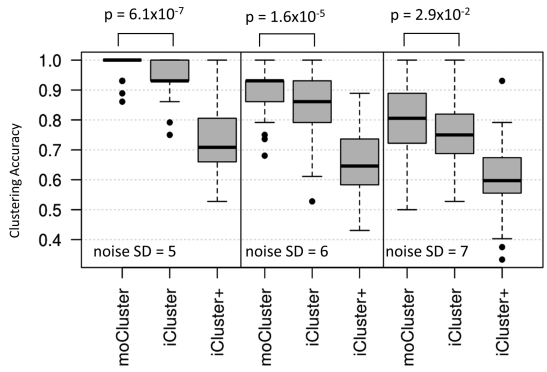



**Figure 3.** Comparison of moCluster, iCluster, and iCluster+ using simulated data based on real biological data (simData2). The three methods were applied to 100 simulated data sets, and their performance was measured by clustering accuracy (see main text). Using different defined experimental error levels (noise SD = 5, 6, and 7, respectively), moCluster consistently outperforms the other methods. *p* values were calculated using the Wilcoxon rank-sum test.

In addition, we compared the effects of the different approaches for calculating JLVs, including CPCA, GCCA, and MCIA. The results suggest that the performance of CPCA and MCIA is comparable and that both of them significantly outperform the GCCA approach, which is the only one that results in nonorthogonal JLVs (Figure S3). However, the deflation strategy used by MCIA determines that coefficient vectors will result in strong interference with the JLVs. Hence, the CPCA approach appears to be the most suitable method for clustering purposes.

## Application of moCluster to Molecular Profiling Data of the NCI-60 Cell Line Panel

Simulated data are valuable for initial benchmark tests of an algorithm, but they tend to oversimplify a biological system and may therefore not reflect its true complexity. Hence, we applied moCluster to cluster mRNA and proteomics data of the NCI-60 cell line panel.[10] The NCI-60 cell line panel consists of 60 cancer cell lines originating from nine different tissues (skin, breast, lung, ovary, prostate, blood, central nervous system, colon, and kidney) and has been extensively studied, for


DOI: 10.1021/acs.jproteome.5b00824
J. Proteome Res. 2016, 15, 755−765

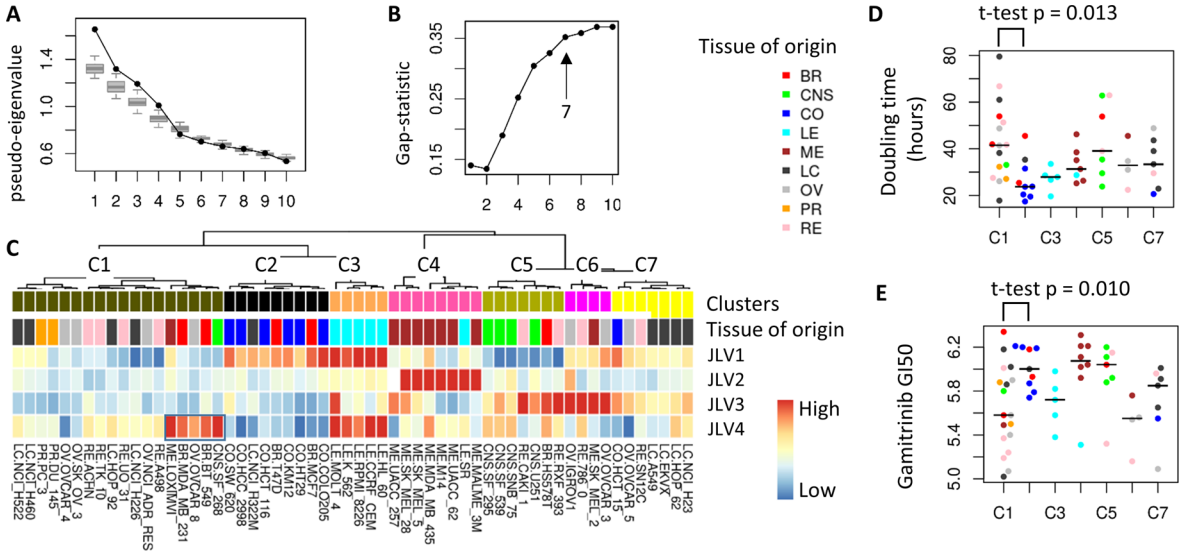



**Figure 4.** Application of moCluster to NCI-60 mRNA and proteomics data. (A) Permutation test to determine the number of latent variables that should be included in the clustering. The light gray boxplot shows the eigenvalues for a 30 permutation test. The first four latent variables represent the concordant structure in the two data sets and are significantly higher than the others. Therefore, we selected the first four latent variables for our analysis. (B) Gap statistic for the clustering of the top four latent variables. From two clusters onward, the gap statistic continuously increased until it reached saturation at approximately seven clusters. (C) Clustering of the top four latent variables using hierarchical clustering. Color of the bars indicates the subtype, tissue of origin, and latent variable. (D) Cell doubling time of different clusters. (E) Different sensitivity to the HSP90 inhibitor Gamitrinib across different subtypes of cancer cell lines.

example, in the context of drug sensitivity.[11] After filtering (see Methods for details), the microarray-based transcriptomic and mass spectrometry-based proteomic data consist of 11 826 and 8069 genes/proteins, respectively. We had to exclude two of the cell lines because of data availability, resulting in 58 cell lines (see Methods). Due to the inherent complexity of cancer biology, neither the transcriptomic nor the proteomic data necessarily distinguished all of the cell lines with respect to their tissue origin. However, the melanoma cell lines in the transcriptomic data and the leukemia cell lines in the proteomic data are significantly different compared to others due to the expression of genes/proteins related to melanogenesis and the immune response (Figure S4).[1] Therefore, our evaluation of moCluster focused on distinguishing these two types of cell lines from the others. In the latent variable space defined by moCluster (using transcriptomic and proteomic data), we observed that most of the melanoma and leukemia cell lines cluster according to their tissues of origin and are clearly separated from other cell lines (Figure S5). Due to the nondeterministic property of iCluster+, we executed this algorithm 10 times. In none of even two cases were the results the same, and two runs entirely failed to distinguish the melanoma cell lines from others (Figure S5). In addition, the computation time to run iCluster+ was 100 to 1000 times longer in comparison to that for moCluster (Table S1). Next, we combined the permutation test and scree plot to determine the number of latent variables that should be included in the analysis. The permutation test showed that the top four JLVs account for significantly correlated structures. The eigenvalue scree plot did not show a clear elbow point, and the top four JLVs were significantly higher than the rest (Figure 4A). Therefore, we have included four JLVs in our analysis.

In order to derive a model based on sparse feature coefficients, we evaluated 10%, 20%, and 40% nonsparse feature coefficients in each of the data sets. The latent variables with as low as 10% sparse coefficients are well correlated with the ones from nonsparse coefficients (Figure S6). Therefore, we chose 10% nonsparse coefficients (the most parsimonious model) and hypothesized that JLVs with sparse coefficients capture essentially the same biological information as their nonsparse counterparts. Then, we applied hierarchical clustering to the four JLVs. In order to determine the optimal number of clusters, the gap statistic was employed.[12] For a given cluster model, the gap statistic calculates the difference (gap) between the within-cluster dispersion of models being evaluated and the model derived from a proper null reference distribution. The within-cluster dispersion is measured as the pooled within-cluster sum of square to the center of each cluster. A high gap statistic indicates that a certain cluster model outperforms a random model generated from the null distribution. In our analysis of the NCI-60 panel, the gap statistic kept increasing until a number of 10 clusters was reached. However, we observed that from seven clusters onward the increase in the gap statistic became moderate (Figure 4B). Therefore, we selected a seven-cluster model (C1−C7), resulting in a good compromise between accuracy and parsimony. The seven resulting subtypes are shown in Figure 4C, with most of the leukemia and melanoma cell lines converging according to their tissues of origin. The heatmap of latent variables depicts that leukemia cell lines have high values of the first latent variable, whereas melanoma cell lines show high values in the second JLV.

In order to better understand the underlying biological processes driving the subtypes, we selected genes and proteins with nonzero coefficients and passed them to a gene set over-representation analysis (DAVID functional annotation).[13] This analysis revealed that genes and proteins positively associated with the first latent variable are related to DNA replication, lymphocyte/T cell activation, and DNA repair processes (Table S2; a full list of enriched gene sets can be found in Table S3). The over-representation of lymphocyte/T cell activation is in concordance with leukemia cell lines showing high values of



DOI: 10.1021/acs.jproteome.5b00824
J. Proteome Res. 2016, 15, 755−765

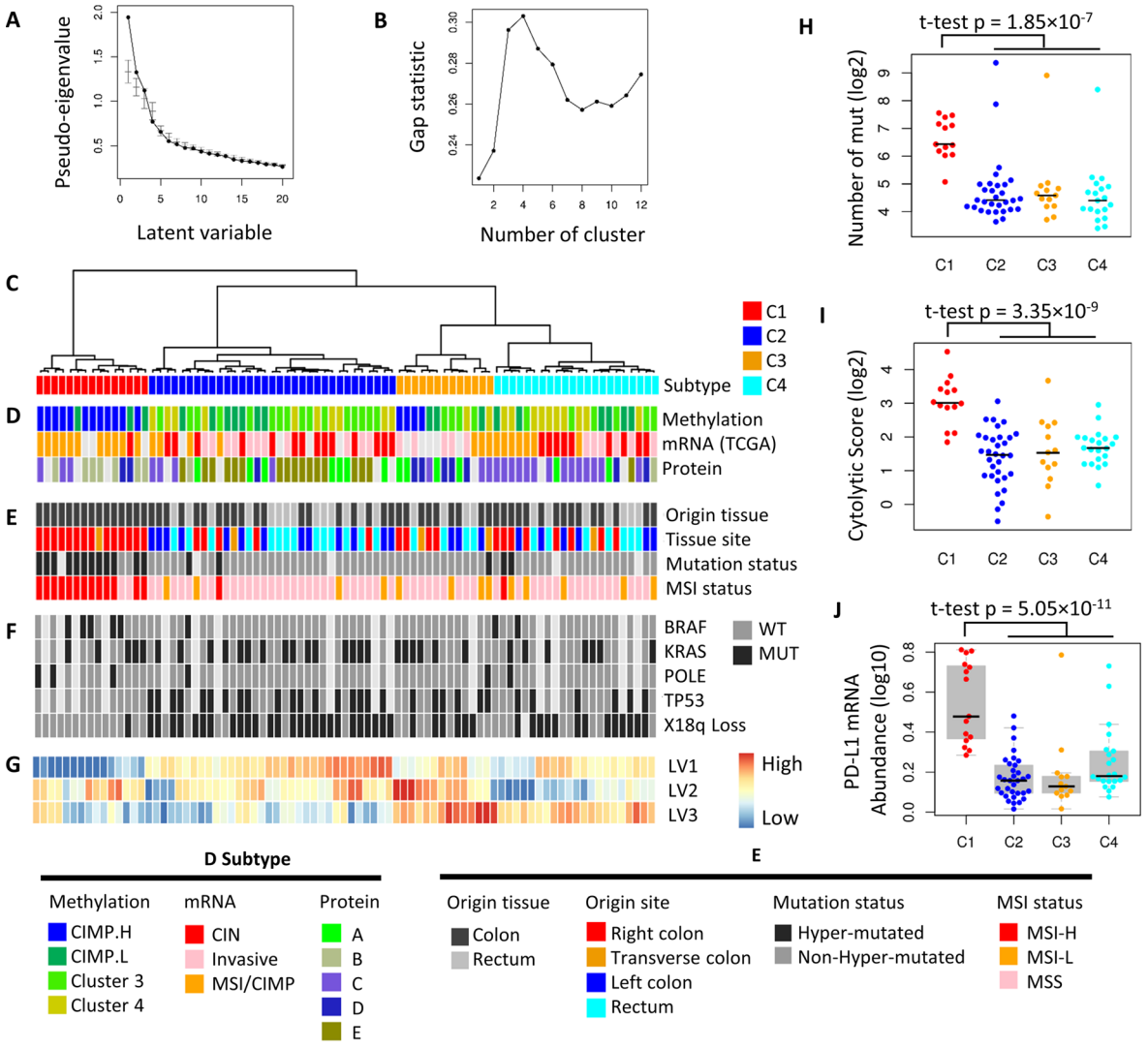



**Figure 5.** Application of moCluster to the TCGA colorectal cancer data set. (A) Variance of the top three latent variables is significantly higher than the others. Two of them represent the concordant structure across the three data sets, as suggested by the permutation test (the error bars represent the 95% confidence interval). (B) Gap statistic with respect to 1 to 12 subtypes indicates that a four-subtype model is the optimal choice. (C) Cluster assignment of patients. (D) Comparison of the integrative subtype model with other subtype models derived from analyzing individual data sets. (E) Comparison of the integrative subtype model with clinical information. (F) Mutation patterns of colorectal cancer related genes in different subtypes. (G) Heatmap showing the latent variable expression pattern. (H) Different distribution of cytolytic scores across subtypes. Subtype C1 is significantly more cytolytic than the others. (I) Number of mutations of patients in different subtypes. Subtype C1 harbors significantly more mutations than the others. (J) Different PDL1 expression in different subtypes.

this latent variable. It is noteworthy that the leukemia cell line SR clustered together with the melanoma cell lines due to its protein but not its mRNA expression profile. This might indicate a mislabeling of the cell line in the proteomics data. We also observed that other clusters with high values of the first JLV, including C2 (mainly comprising colon cell lines) and C3 (leukemia cell lines), possess the shortest doubling times (ANOVA $p$ value = 0.030; Figure 4D). This is in agreement with the over-representation of DNA replication genes and proteins in this latent variable.

Cell adhesion and cellular motions are associated with negative values of latent variables 1 and 2 as well as positive values of JLV 3. These three JLVs mainly drive one big cluster, C1, and two small clusters, C5 and C6. This is in concordance with the epithelial nature of these subtypes. High values of the second latent variable, a defining characteristic of melanoma cell lines, are strongly associated with genes and proteins related to melanogenesis and pigmentation function. One of the melanoma cell lines, LOXIMVI, known to lack the proteins/genes for melanogenesis was therefore correctly clustered away from the other melanoma cells and was associated with the nonsmall cell lung cancer cells, both being epithelial cells.

The mRNAs associated with high values of the fourth JLV (JLV4) are also enriched in T cell activation and lymphocyte activation. Apart from the leukemia subtypes, high values of this latent variable additionally defined a subgroup in the C1 subtype. This subgroup includes two claudin low breast cancer cell lines, BT549 and MDAMB231. It was shown before that claudin low cancers have a higher expression of genes related to T cell, B cell, and granulocyte function.[14] Interestingly, three other cell lines, the CNS cell line SF268, the ovarian cell line OVCR 8, and the melanoma cell line LOXIMVI, were also found in this this subgroup, which implies that these cancer cell


DOI: 10.1021/acs.jproteome.5b00824
J. Proteome Res. 2016, 15, 755−765

lines may possess characteristics comparable to the claudin low breast cancer cell lines (JLV4 in Figure 4C).

Other functions enriched by selected genes and proteins (gene/proteins with nonzero coefficients) are closely related to GTPase regulation (Tables S2 and S3) and mitochondrial activity, including the electron transport chain, protein oxidative phosphorylation (protein data with negative values of JLV3), and oxidative reduction (proteins on negative end of JLV4). Therefore, this cancer subtype may show a distinct drug sensitivity profile for drugs targeting the mitochondrion.[15] In an attempt to validate this hypothesis, we compared the growth inhibition effect ($GI_{50}$) of Gamitrinib, a mitochondrial HSP90 ATPase inhibitor, among different subtypes (ANOVA $p$ value = 0.010; Figure 4E). This analysis suggested that different subtypes indeed show different sensitivity to this drug, with subtypes C2, C4, and C5 being more resistant to this compound. In summary, moCluster successfully captures the important joint pattern defined by both proteomic and transcriptomic data. The biological interpretation of the clusters was facilitated by exploring the features associated with latent variables.

## Application of moCluster to Molecular Profiling Data of Colorectal Cancer Patients

The Cancer Genome Atlas (TCGA) and Clinical Proteomic Tumor Analysis Consortium (CPTAC) recently published multidimensional, genome-scale and proteome-scale analyses on colon and rectum carcinoma.[16] Several different subtype models were proposed by clustering individual omics data from methylation, transcriptomic, and proteomic studies: four subtypes were discovered in the methylation data set, including two subtypes with elevated methylation, designated CIMP high and CIMP low, and the non-CIMP clusters, cluster 3 and cluster 4;[16a] three transcriptomic subtypes were designated microsatellite instability/CpG island methylater phenotype (MSI/CIMP), invasive, and chromosomal instability (CIN);[16a] Zhang et al. reported five proteomic subtypes, designated subtypes A−E.[16b] However, the proteomics study observed a limited correlation between mRNA and protein levels.[16b] Therefore, we asked whether colorectal cancer subtypes can be better represented by an integrative clustering based on the three types of data. For this purpose, we applied the moCluster algorithm to a subset of 83 tumors for which all three types of data were available.

A permutation test suggested that the first two JLVs represent a significant coherent structure among the data sets. In conjunction with the eigenvalue plot, we decided to use three JLVs for clustering (Figure 5A). Ten percent nonzero coefficient (the same selection procedure as for the NCI-60 example discussed above) in each data set was retained. There was a good correlation between sparse and nonsparse JLVs (Figure S7). Then, we clustered the top three JLVs using hierarchical clustering (Euclidean distance with Ward's method) and retained four robust integrative subtypes based on the gap statistic (Figure 5B), denoted C1−C4, consisting of 15, 33, 13, and 22 cases, respectively (Figure 5C). Silhouette analysis confirmed the robustness of this model, i.e., except for one patient in C2 and one in C4, most of the silhouette widths of patients are positive (Figure S8). In addition, to evaluate how the integrative clustering is different from clustering individual data sets, a similar clustering approach based on the PCs of individual data sets was applied (Figure S9). The results suggest six subtypes in the methylation data and four subtypes in the mRNA data, whereas no obvious clustering structure was found for the proteomics data (Figure S9). Although some individual subtypes are significantly associated with the integrative subtypes (e.g., methylation subtype 3 to C1, transcriptomics subtype 1 to C2 and C4; Table S4), the integrative subtype is still a different subtype model. For example, mRNA subtype 1 is divided into two integrative subtypes, C2 and C4, which could be obtained only from the heterogeneous methylation and/or proteome patterns of patients in this subtype.

Next, we tested the association between our integrated subtype model with other established subtype models using Fisher's exact test (Figure 5D and Table S5). Integrated subtype C1 is significantly enriched with patients from proteomics subtype B, methylation subtype CIMP.H, and MSI/CIMP patients. Concordantly, we observed high MSI and methylation status in these patients, and 14 out of 15 originate from the right colon (Figure 5E and Table S5). In addition, this subtype is significantly associated with the absence of the P53 mutation and the absence of chromosome 18q and is moderately enriched in patients with BRAF mutations (Figure 5F and Table S5). The independent discovery of this subtype in several different studies suggests that this subtype is very different from the other subtypes on several regulatory levels. In our analysis, the dendrogram suggested that this subtype is most distinct from the remaining subtypes (Figure 5C), which is specifically characterized by low values in the first JLV (Figure 5G).

Of note, the other three integrative subtypes were not discovered in previous studies. In particular, the mRNA subtype chromosome instability (CIN), a well-accepted genetic property of colorectal cancer, can be subdivided into subtypes C2 and C4. This result implies that different mechanisms of oncogenesis may be present in tumors of the mRNA CIN subtype. The C2 subtype also included most of the proteome subtype E, which is characterized by HNF4A amplification and, as a consequence, by higher protein levels of HNF4$\alpha$.[16b] However, we still consider C2 to be a new subtype because this subtype also contains 13 patients from other proteomic subtypes. Furthermore, there are only weak associations present between proteomic subtype D and C3 as well as proteomic subtype C and the integrative cluster C4 (Table S5).

Colorectal cancers are generally divided into two main groups: microsatellite unstable (MSI) or microsatellite stable (MSS) but chromosomally unstable tumors.[16a] MSI tumors are also characterized as being primarily located in the right colon, harboring the CpG island methylator phenotype (CIMP) and being hypermutated.[16a] In our analysis, we observed that C1 patients have a higher mutational frequency ($t$-test $p$ value = $1.85 \times 10^{-7}$; Figure 5H). Therefore, this subtype corresponds to our integrative subtype C1. An enrichment analysis of the associated features (features with negative coefficients in LV 1) suggested that low values of the first latent variable are associated with immune-related genes and proteins (Tables S6 and S7). This implies that this subtype may also represent an immune infiltrated subtype. In order to quantify the degree of immune activation in the samples, we used the cytolytic activation score from Rooney and colleagues.[17] The result confirmed that C1 tumors have the highest degree of immune activation ($t$-test $p$ value = $3.35 \times 10^{-9}$; Figure 5I). An increasing mutational load is thought to increase the number of cancer associated antigens presented by tumor cells, thus eliciting an enhanced immune response. These cancers are


DOI: 10.1021/acs.jproteome.5b00824
J. Proteome Res. 2016, 15, 755−765

thought to evade immune surveillance and eradication through the expression of PDL1, and, interestingly, cancers with high levels of mutational heterogeneity have responded well to anti-PD-1 therapy in early clinical trials.[18] Specifically, Llosa et al. reported that a specific immune microenvironment is associated with MSI colorectal cancer and suggested five immune checkpoint genes, including PD-1, PD-L1, CTLA-4, LAG-3, and IDO, as potential drug targets.[19] Of note, all of these checkpoints are significantly (*t*-test *p* values range from $10^{-3}$ to $10^{-11}$) elevated in our integrative subtype C1 (MSI and immune activation subtype; Figures 5J and S10). This is particularly interesting since there are several colorectal cancer drugs targeting immune checkpoints in early clinical trials, including anti-PD-1 antibodies Nivolumab (NCT02060188), MEDI0680 (NCT02118337; NCT02013804), and pembrolizumab (NCT01876511); anti-CTLA-4 antibodies ipilimumab (NCT02060188) and tremelimumab (NCT02205333); anti-CD27 antibody varlilumab (NCT01460134); and a LAG-3 antibody BMS-986016 (NCT01968109). Therefore, the C1 subtype identified by our analysis may represent a subset of colorectal cancers, which might be well susceptible to drugs targeting these immune checkpoint genes. Conversely, the other three subtypes may be less sensitive to this treatment.

Subtypes C2−C4 were not described before; hence, the integrative subtype model provides a new basis to study the mechanisms driving these colorectal cancers. In particular, C2 subtype tumors are characterized by negative weight on the third JLV, whereas subtypes C3 and C4 are distinct on the second JLV. To understand the related functions, we performed enrichment analysis of the features associated with the second and third latent variables. We observed that, although a limited correlation between mRNA and protein was reported,[16b] a relatively good correlation was observed between them on the gene set level, namely, the selected nonzero weight mRNAs and proteins seem to be enriched in the same gene sets (Tables S6 and S7). The result suggested that C2 tumors have elevated ribosome biogenesis activity, and the associated gene sets actually include molecular functions related to ribosome biogenesis and RNA processing. An increased demand for ribosome biogenesis has been associated with tumorigenesis and an increased risk of neoplastic transformation.[20] The C3 subtype has high values of the second JLV; the associated genes include collagens, integrins, and cadherins, which are functionally involved in cell adhesion and immune related processes (Table S7). These molecules play a role in the attachment of malignant cells in their original site, whereas the down-regulation of these genes may support the metastasis of cancer cells to foreign tissues in colon cancer.[21] Therefore, C2 might be a subtype with more advanced neoplastic transformation, whereas C3 may represent a subtype with a more epithelial phenotype and less metastatic potential. This analysis provides a hypothesis that may be further tested in the future.

## CONCLUSIONS

In this study, we present a new method, moCluster, to identify the joint molecular patterns in multiple omics data. We show that the algorithm can generate clustering models that cannot be obtained by single data analysis alone. Harnessing the benefits of multitable multivariate analysis, moCluster identifies robust latent variables and runs hundreds of times faster than alternatives such as the iCluster algorithm. At the same time, a sparse operator is incorporated, enabling the selection of important features associated with each JLV. This greatly facilitates the biological interpretation of latent variables and therefore aids in the identification of novel biological hypotheses, which can be subsequently tested by further experiments.

However, moCluster, like other computational methods, should not be used blindly. Our previous work in this area (the MICA approach) showed that the concordance between transcriptomic and proteomic data is increased when filtering out missing values in the proteomics data.[1] Therefore, differences in results between omics data could result from technical artifacts (e.g., missing values), which may lead to further artifacts in joint patterns across data sets. Thus, careful quality control of each individual data set is required before any integrative analysis is performed. Furthermore, this study mainly considered data that can be modeled using normal distributions (such as log-transformed microarray normalized intensity, RPKM, and normalized intensity protein expression data). Appling moCluster to other types of data requires different normalization steps. For example, count data encountered in RNA-Seq (read count) or mass spectrometry (spectrum count) may be converted to a chi-squared matrix, as in correspondence analysis (CA) or nonsymmetric correspondence analysis (NSCA).[1,22] It worth noting that, because of their descriptive nature, the normalization methods in CA or NSCA could also be used for other data types.[23] Hence, in some cases, such normalization methods may lead to better clustering results.[24]

In summary, we believe that moCluster will greatly facilitate the integrative analysis of data, as there are increasingly more studies generating multiple omics data for the same set of samples. The moCluster algorithm is available in the Bioconductor package "mogsa" (version 1.1.5 or newer).

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.5b00824.

iCluster+ replicate runs; iCluster results with up to 50 iterations; comparison of different extensions of PCA; PCA of transcriptomic and proteomic data from NCI-60 cell lines; JLV space defined by moCluster and iCluster+; correlation between sparse and non-sparse solutions for latent variables; silhouette plot for the integrative subtype model; evaluation of the clustering structure of each dataset from the colorectal panel; mRNA expression of four immune-related genes; computation time of different algorithms; enrichment analysis of variables associated with each latent variable; comparison of integrative subtypes with subtypes defined by individual datasets; associations between subtype and other subtype/clinical factors (PDF)

Over-representation analysis by DAVID functional annotation for the NCI-60 datasets (XLSX)

Over-representation analysis by DAVID functional annotation for the TCGA colorectal datasets (XLSX)

## AUTHOR INFORMATION

### Corresponding Author

*E-mail: kuster@tum.de.



DOI: 10.1021/acs.jproteome.5b00824
J. Proteome Res. 2016, 15, 755−765

## Notes

The authors declare no competing financial interest.

## REFERENCES


(1) Meng, C.; Kuster, B.; Culhane, A. C.; Gholami, A. M. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinf.* **2014**, *15*, 162.

(2) Hoadley, K. A.; Yau, C.; Wolf, D. M.; Cherniack, A. D.; Tamborero, D.; Ng, S.; Leiserson, M. D.; Niu, B.; McLellan, M. D.; Uzunangelov, V.; Zhang, J.; Kandoth, C.; Akbani, R.; Shen, H.; Omberg, L.; Chu, A.; Margolin, A. A.; Van't Veer, L. J.; Lopez-Bigas, N.; Laird, P. W.; Raphael, B. J.; Ding, L.; Robertson, A. G.; Byers, L. A.; Mills, G. B.; Weinstein, J. N.; Van Waes, C.; Chen, Z.; Collisson, E. A.; Cancer Genome Atlas Research Network; Benz, C. C.; Perou, C. M.; Stuart, J. M. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **2014**, *158* (4), 929−44.

(3) Monti, S.; Tamayo, P.; Mesirov, J.; Golub, T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Mach. Learn.* **2003**, *52* (1−2), 29.

(4) Lee, H.; Kong, S. W.; Park, P. J. Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes. *Bioinformatics* **2008**, *24* (7), 889−96.

(5) Shen, R.; Olshen, A. B.; Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **2009**, *25* (22), 2906−12.

(6) Curtis, C.; Shah, S. P.; Chin, S. F.; Turashvili, G.; Rueda, O. M.; Dunning, M. J.; Speed, D.; Lynch, A. G.; Samarajiwa, S.; Yuan, Y.; Graf, S.; Ha, G.; Haffari, G.; Bashashati, A.; Russell, R.; McKinney, S.; Group, M.; Langerod, A.; Green, A.; Provenzano, E.; Wishart, G.; Pinder, S.; Watson, P.; Markowetz, F.; Murphy, L.; Ellis, I.; Purushotham, A.; Borresen-Dale, A. L.; Brenton, J. D.; Tavare, S.; Caldas, C.; Aparicio, S. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **2012**, *486* (7403), 346−52.

(7) Mo, Q.; Wang, S.; Seshan, V. E.; Olshen, A. B.; Schultz, N.; Sander, C.; Powers, R. S.; Ladanyi, M.; Shen, R. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (11), 4245−50.

(8) Westerhuis, J. A.; Kourti, T.; Macgregor, J. F. Analysis of multiblock and hierarchical PCA and PLS moels. *J. Chemom.* **1998**, *12* (5), 301−21.

(9) Hassani, S.; Hanafi, M.; Qannari, E. M.; Kohler, A. Deflation strategies for multi-block principal component analysis revisited. *Chemom. Intell. Lab. Syst.* **2013**, *120*, 154−68.

(10) Moghaddas Gholami, A.; Hahne, H.; Wu, Z.; Auer, F. J.; Meng, C.; Wilhelm, M.; Kuster, B. Global proteome analysis of the NCI-60 cell line panel. *Cell Rep.* **2013**, *4* (3), 609−20.

(11) Reinhold, W. C.; Sunshine, M.; Liu, H.; Varma, S.; Kohn, K. W.; Morris, J.; Doroshow, J.; Pommier, Y. CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Res.* **2012**, *72* (14), 3499−511.

(12) Tibshirani, R.; Walther, G.; Hastie, T. Estimating the number of data clusters via the Gap statistic. *Journal of the Royal Statistical Society B* **2001**, *63*, 411−23.

(13) Huang, D. W.; Sherman, B. T.; Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **2008**, *4* (1), 44−57.

(14) Sabatier, R.; Finetti, P.; Guille, A.; Adelaide, J.; Chaffanet, M.; Viens, P.; Birnbaum, D.; Bertucci, F. Claudin-low breast cancers: clinical, pathological, molecular and prognostic characterization. *Mol. Cancer* **2014**, *13*, 228.

(15) Fulda, S.; Galluzzi, L.; Kroemer, G. Targeting mitochondria for cancer therapy. *Nat. Rev. Drug Discovery* **2010**, *9* (6), 447−64.

(16) (a) Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **2012**, *487* (7407), 330−7. (b) Zhang, B.; Wang, J.; Wang, X.; Zhu, J.; Liu, Q.; Shi, Z.; Chambers, M. C.; Zimmerman, L. J.; Shaddox, K. F.; Kim, S.; Davies, S. R.; Wang, S.; Wang, P.; Kinsinger, C. R.; Rivers, R. C.; Rodriguez, H.; Townsend, R. R.; Ellis, M. J.; Carr, S. A.; Tabb, D. L.; Coffey, R. J.; Slebos, R. J.; Liebler, D. C.; Nci, C. Proteogenomic characterization of human colon and rectal cancer. *Nature* **2014**, *513* (7518), 382−7.

(17) Rooney, M. S.; Shukla, S. A.; Wu, C. J.; Getz, G.; Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **2015**, *160* (1−2), 48−61.

(18) (a) Champiat, S.; Ferte, C.; Lebel-Binay, S.; Eggermont, A.; Soria, J. C. Exomics and immunogenics: Bridging mutational load and immune checkpoints efficacy. *Oncoimmunology* **2014**, *3* (1), e27817. (b) Powles, T.; Eder, J. P.; Fine, G. D.; Braiteh, F. S.; Loriot, Y.; Cruz, C.; Bellmunt, J.; Burris, H. A.; Petrylak, D. P.; Teng, S. L.; Shen, X.; Boyd, Z.; Hegde, P. S.; Chen, D. S.; Vogelzang, N. J. MPDL3280A (anti-PD-L1) treatment leads to clinical activity in metastatic bladder cancer. *Nature* **2014**, *515* (7528), 558−62.

(19) Llosa, N. J.; Cruise, M.; Tam, A.; Wicks, E. C.; Hechenbleikner, E. M.; Taube, J. M.; Blosser, R. L.; Fan, H.; Wang, H.; Luber, B. S.; Zhang, M.; Papadopoulos, N.; Kinzler, K. W.; Vogelstein, B.; Sears, C. L.; Anders, R. A.; Pardoll, D. M.; Housseau, F. The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints. *Cancer Discovery* **2015**, *5* (1), 43−51.

(20) Montanaro, L.; Trere, D.; Derenzini, M. Nucleolus, ribosomes, and cancer. *Am. J. Pathol.* **2008**, *173* (2), 301−10.

(21) Paschos, K. A.; Canovas, D.; Bird, N. C. The role of cell adhesion molecules in the progression of colorectal cancer and the development of liver metastasis. *Cell. Signalling* **2009**, *21* (5), 665−74.

(22) Fellenberg, K.; Hauser, N. C.; Brors, B.; Neutzner, A.; Hoheisel, J. D.; Vingron, M. Correspondence analysis applied to microarray data. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98* (19), 10781−6.

(23) Greenacre, M. *Correspondence Analysis in Practice* **2007**, DOI: 10.1201/9781420011234.

(24) Wouters, L.; Gohlmann, H. W.; Bijnens, L.; Kass, S. U.; Molenberghs, G.; Lewi, P. J. Graphical exploration of gene expression data: a comparative study of three multivariate methods. *Biometrics* **2003**, *59* (4), 1131−9.

(25) Hanafi, M.; Kohlerb, A.; Qannari, E.-M. Connections between multiple co-inertia analysis and consensus principal component analysis. *Chemom. Intell. Lab. Syst.* **2011**, *106* (1), 37.

(26) Murtagh, F.; Legendre, P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J. Classif.* **2014**, *31* (274), 95.

(27) Schwanhausser, B.; Busse, D.; Li, N.; Dittmar, G.; Schuchhardt, J.; Wolf, J.; Chen, W.; Selbach, M. Global quantification of mammalian gene expression control. *Nature* **2011**, *473* (7347), 337−42.

(28) Cerami, E.; Gao, J.; Dogrusoz, U.; Gross, B. E.; Sumer, S. O.; Aksoy, B. A.; Jacobsen, A.; Byrne, C. J.; Heuer, M. L.; Larsson, E.; Antipin, Y.; Reva, B.; Goldberg, A. P.; Sander, C.; Schultz, N. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery* **2012**, *2* (5), 401−4.

(29) Du, P.; Zhang, X.; Huang, C. C.; Jafari, N.; Kibbe, W. A.; Hou, L.; Lin, S. M. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinf.* **2010**, *11*, 587.

(30) Zhu, Y.; Qiu, P.; Ji, Y. TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nat. Methods* **2014**, *11* (6), 599−600.




DOI: 10.1021/acs.jproteome.5b00824
*J. Proteome Res.* 2016, 15, 755−765